공학박사학위논문

Machine Learning Techniques for Short-range Meteorological Forecasts

(단기 기상 예측을 위한 기계 학습 기법)

2020년 2월

서울대학교 대학원

전기·컴퓨터공학부

문 승 현

Machine Learning Techniques for Short-range Meteorological Forecasts

지도교수 문 병 로

이 논문을 공학박사 학위논문으로 제출함.

2019년 11월

서울대학교 대학원

전기·컴퓨터공학부

문 승 현

문승현의 공학박사 학위논문을 인준함.

2019년 12월

위 원 장 신 영 길 (인)

부위원장 문 병 로 (인)

위 원 이 광 근 (인)

위 원 김 용 혁 (인)

위 원 권 영 근 (인)

Abstract

Machine Learning Techniques for Short-range Meteorological Forecasts

Seung-Hyun Moon

Department of Electrical Engineering & Computer Science

The Graduate School
Seoul National University

Machine learning is the study of artificial intelligence that automatically generates programs from data. It is distinguished from conventional programming, which needs to write a series of specific instructions directly to perform a specific task. Machine learning is preferred when it is difficult to develop an effective algorithm for given tasks such as natural language processing or computer vision.

Traditionally, numerical weather prediction (NWP) has been a prevailing method to forecast weather. The NWP predicts future weather through simulations using mathematical models based on current weather conditions. However, the NWP has some problems: errors in the current observations are amplified as simulation proceeds; spatial and temporal resolutions are limited; and there is a spin-up problem, in which initial forecasts are unreliable while the model attempts to stabilize. An alternative approach is needed to complement NWP on small spatial and temporal scales. Therefore, we propose short-range weather forecast models that employ machine learning techniques appropriate for a given forecasting problem.

First, we introduce dimensionality reduction techniques to construct effective forecasting models with high-dimensional input data. As the dimension of input data increases, the amount of time or memory required by machine learning techniques can increase significantly. This phenomenon is referred to as the curse of dimensionality, which can be

alleviated by dimensionality reduction techniques. Dimensionality reduction techniques include feature selection and feature extraction. Feature selection selects a subset of input variables, while feature extraction projects high-dimensional features to a lower dimensional space. The details of correlation-based feature selection, and principal component analysis (PCA) which is a representative feature extraction are provided. We then propose a scheme for precipitation type forecast as an example of meteorological forecasting using dimensionality reduction techniques. This scheme takes 93 meteorological variables as input, and uses feature selection to assemble an effective subset of input variables. Multinomial logistic regression is used to classify precipitation as rain, snow, or sleet. This scheme achieved predictions which are 13 % more accurate than the original forecasts, and feature selection improved the accuracy to a statistically significant level.

Second, we present sampling techniques that help predict rare meteorological events. Machine learning algorithms tend to sacrifice performance on rare instances to overall performance, which is referred to as class imbalance problem. To resolve this problem, undersampling reduces the number of common instances. As an example of meteorological forecasting using undersampling, we propose a scheme for lightning forecast. Meteorological variables from European Centre for Medium-range Weather Forecasts provide the input to our scheme, in which an undersampling is used to alleviate the class imbalance problem, and SVMs are used to forecast lightning activities within a particular location and time interval. When the scheme was trained with the original input data, it could not predict any lightning. After undersampling, however, the scheme successfully detected about 38 % of the lightning strikes.

Finally, we propose a selective discretization technique that automatically selects and discretizes suitable variables for discretization. Discretization is a preprocessing technique that converts continuous variables into categorical ones. Conventional discretization techniques apply discretization to all variables, which may lead to significant information loss. The selective discretization minimizes information loss by discretizing only variables that have nonlinear relationship with the dependent variable. We suggest a scheme for heavy rainfall forecast as an example of meteorological forecasting using the selective discretization. This scheme takes input from automatic weather stations, and predicts whether or not the heavy rain criterion will be met within the next three hours. The input variables are preprocessed to have a compressed yet efficient representation through the selective discretization and

PCA. Logistic regression uses the preprocessed data to predict whether or not the heavy rain condition will be satisfied. The selective discretization selectively discretized continuous variables such as date and temperature, contributing to the improvement of predictive performance to a statistically significant level.

We present effective machine learning techniques for short-range weather forecast, and provide case studies that apply machine learning to precipitation type forecast, lightning forecast, and heavy rainfall forecast. We combine appropriate techniques to solve each forecasting problem effectively, and the resulting prediction models were good enough to be used for operational forecasting system.

Keywords : Machine learning, meteorological forecast, dimensionality reduction, undersampling, discretization.

Student Number : 2004-23567

Contents

List of Figures

List of Tables

Chapter 1

Introduction

People have long predicted the weather, and techniques for weather forecasting have changed with the times. Currently, the most prevalent method is numerical weather prediction (NWP), which simulates the atmosphere and oceans to predict the weather based on current weather conditions. However, the NWP has limited spatial and temporal resolution, and there is a so-called spin-up problem, which makes initial forecasts unreliable until the forecasting model is stabilized.

Recently, machine learning has been spotlighted in various fields such as image recognition, natural language processing, and recommendation systems. In this thesis, we propose prediction models that forecast meteorological events well in small spatial and temporal scale by employing machine learning techniques.

1.1 Machine Learning

Machine learning is a type of statistical techniques that automatically generates programs from data without explicit programming. Based on given data, machine learning algorithms build a mathematical model that performs a specific task through pattern recognition and

inference. Normally, data sets in machine learning are divided into two disjoint subsets: a training set is used for constructing a model; and a test set is for evaluating the model.

Machine learning on labeled data is called supervised learning, which finds a function that maps an instance to a label based on a given training set, and uses the function to make predictions on unseen data. Most of the techniques used in this thesis are supervised learning. Forecasting models are trained to predict heavy rainfall, precipitation types, or lightning activities.

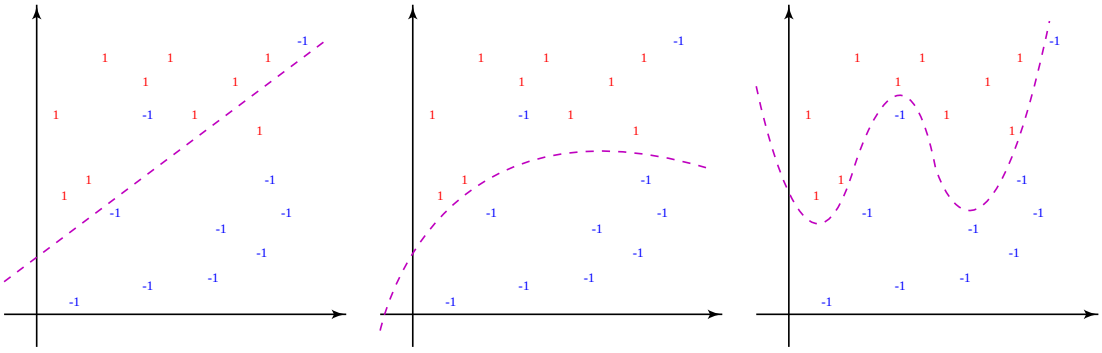
Unlike supervised learning, where each instance has a label to a certain class, machine learning on unlabeled data is referred to as unsupervised learning. While supervised learning is trained to predict desired output labels, unsupervised learning is used to discover effective representation of the data, or cluster the data into groups. In this thesis, principal component analysis, one of the unsupervised learning, is used to reduce the dimension of input data.

1.1.1 Data Preprocessing

Instances in machine learning are described by *features*. For example, when creating a model that determines a person's creditworthiness, instances having features such as age, assets, occupation, and salary will be used. Data preprocessing is a technique that converts raw data into useful information from a machine learning perspective.

The phrase "garbage in, garbage out" emphasizes the importance of data preprocessing in machine learning and data mining. If training data contains irrelevant, redundant, or inefficient forms of information, it will degrade the performance of learning algorithms.

Unfortunately, there is no effective preprocessing method for all problems, so we have to find a suitable method for each problem by trial and error. In this thesis, we describe several preprocessing methods that might come in handy when dealing with large amount of raw meteorological data. Specifically, we introduce dimensionality reduction techniques that reduce the dimension of high-dimensional meteorological data, sampling techniques



(a) Linear function (b) Quadratic function (c) High-dimensional function

Figure 1.1: Examples of decision boundaries for classification

that can help to predict rare meteorological events, and a selective discretization technique that selectively converts numerical variables into categorical variables.

1.1.2 Classification

In machine learning, classification refers to classifying instances into one of two or more classes. In particular, classifying instances into one of two classes is called binary classification. In this thesis, heavy rainfall prediction and lightning forecast belong to binary classification, while precipitation type prediction belong to multiclass (three or more classes) classification.

A classifier is a mathematical model for performing classification, and it is important to choose an appropriate classifier for a given problem. Figure 1.1 shows the examples of decision boundaries for a classifier. A more complicated function can correctly classify more training instances, but it can also have slower learning process, and cause overfitting which hinders generalization. In this thesis, we forecast meteorological events through logistic regression and support vector machines (SVMs). Logistic regression models the probability of a certain class by an S-shaped curve, and SVMs find the maximum-margin hyperplane to separate different classes.

Table 1.1: Meteorological forecasting ranges

Category	Forecast lead time
Nowcasting	0 to 3 hours (Glossary of Meteorology, 2019a)
Very short-range forecasting	0 to 6 hours (Glossary of Meteorology, 2019b)
Short-range forecasting	0 to 72 hours (Nese et al., 2018)
Medium-range forecasting	72 to 240 hours (World Meteorological Organization, 2019)
Extended-range forecasting	10 to 30 days (World Meteorological Organization, 2019)
Long-range forecasting	30 days to two years (World Meteorological Organization, 2019)

1.2 Meteorological Forecasts

Meteorological forecast is the application of science and technology that gathers meteorlogocial elements from multiple locations and then predicts future weather conditions. The atmosphere around the surface is constantly circulating, and there are many variables that affect the atmosphere. There are no known perfect model that can predict future weather, and thus meteorological forecast may not be accurate. Meteorological forecast is classified into several problems listed in Table 1.1, depending on the forecast lead time. This thesis focuses on short-term meteorological forecasts in which conventional numerical weather prediction is severely affected by initial errors.

Recent advances in machine learning have led to innovation in many areas, including image recognition, natural language processing, and automated driving systems. However, it is not easy to find successful applications of machine learning in weather forecasting. We investigate the possibility of machine learning in short-range weather forecasts by applying machine learning techniques to three meteorological events.

1.2.1 Precipitation Types

Precipitation refers to all forms of the condensation of atmospheric water vapor that falls on the ground in the global hydrologic cycle. In winter, rain, snow, and sleet which is a mixture of rain and snow can be observed. Since they have different effects on the ground, accurate prediction of precipitation types is very important. For example, rainfall usually infiltrates into the ground and can contribute to runoff quickly (Mein and Larson, 1973), but snowfall first accumulates on the surface, melts and infiltrates into the soil with a delay (Zhong et al., 2018). The snowfall accumulation increases albedo and alters the surface energy budget (Box et al., 2012). It is also known that the highest accident risk is associated with road slipperiness due to rain or sleet on a frozen road surface (Norrman et al., 2000). Thus, the accurate determination of precipitation types when the surface temperature is near freezing is one of the most serious problems associated with wintertime weather forecasts (Ralph et al., 2005). We aim to improve precipitation type predictions already included in the short-range weather forecasts from European Centre for Medium-Range Weather Forecast and Regional Data Assimilation and Prediction System.

1.2.2 Lightning

Lightning is a good indicator for the detection of severe weather conditions (Schultz et al., 2009) and responsible for tens of thousands of casualties worldwide every year (Holle, 2008). In the United States, lightning is known to kill more people than tornadoes and hurricanes (Curran et al., 2000). In addition, lightning can trigger forest fires (Liu et al., 2016) and aviation accidents (Mäkelä et al., 2013), which cause serious losses of life and property. Therefore, it is necessary to accurately predict lightning activities, one of the most dangerous natural disasters. We predict lightning activities using short-range weather forecasts from European Centre for Medium-Range Weather Forecast.

1.2.3 Heavy Rainfall

Heavy precipitation causes serious losses of life and property, and often triggers natural disasters such as landslides and flash floods. In South Korea, a heavy rain advisory is issued when the expected amount of precipitation is over 70 mm in 6 hours or 110 mm in 12 hours (Korea Meteorological Administration, 2018). Kim et al. (2011) reported that the damage caused by heavy rainfall occurred most frequently at these intensities from 2005 to 2009. Accurate and timely warning information is needed to minimize the damage. Therefore, we construct an early warning system for very short-range heavy rainfall.

1.3 Main Contributions

In this thesis, we introduce machine learning approaches for three meteorological forecasts. The major contributions of this thesis are listed below.

- Dimensionality reduction for high-dimensional meteorological data:
  The weather is a chaotic system, and there are many variables that affect each meteorological element. Applying machine learning to meteorological forecast with high-dimensional data may require significant computing resources or cause overfitting problems due to the curse of dimensionlity. To resolve this issue, feature selection and feature extraction are presented. Feature selection was used for precipitation type prediction that takes input from various meteorological variables, and feature extraction was used for heavy rainfall prediction that uses many correlated variables as input. The dimensionality reduction technique in each forecasting system improved the predictive performance significantly.

- Balancing highly imbalanced meteorological data:
  Meteorological disasters such lightning and typhoons rarely occur, but once they do,

they can cause serious losses of life and property. When machine learning is applied to predict such events, however, it is often impossible to predict these events. Since most machine learning algorithms try to increase overall accuracy, they achieve high accuracy without predicting rare events at all. We suggest undersampling and oversampling to rebalance highly imbalanced meteorological data. A lightning forecasting scheme trained with the original training data did not predict any lightning, but after rebalancing the training data, it was able to predict lightning successfully.

- Selective discretization scheme:

  We propose a selective discretization scheme, which selectively discretizes input variables. Discretization is a preprocessing method that converts continuous variables into discrete ones. Conventional discretization methods discretizes all input variables, but the selective discretization discerns variables for which discretization is appropriate, and discretizes only those variables. It prevents information loss caused by inappropriate discretization of numerical variables. The selective discretization was used for heavy rainfall forecast, and it improved the predictive performance by selectively discretizing numerical variables such as date and temperature.

- Meteorological forecasting models for operational use:

  We suggest a heavy rainfall forecasting model, a precipitation type forecasting model, and a lightning forecasting model. These forecasting models are developed to complement the numerical weather prediction by employing machine learning techniques. They showed promising results that could be used for operational use. Comparative analysis for various techniques will be helpful in constructing other meteorological forecasting models using machine learning techniques.

1.4 Organization

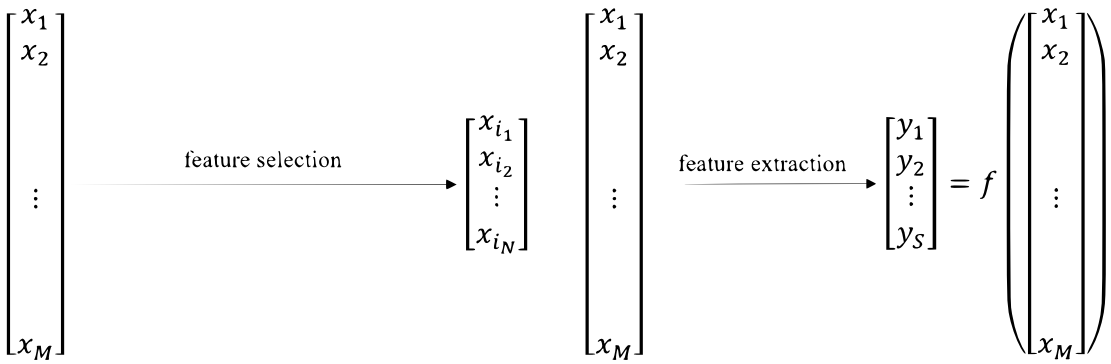
The thesis is organized as follows: in Chapter 2 we introduce dimensionality reduction techniques, and propose a scheme for precipitation type prediction that utilizes the dimensionality reduction techniques. In Chapter 3 we describe sampling techniques that alleviates the class imbalance problem, and propose a machine learning approach to forecast lightning using the sampling methods. In Chapter 4 we suggest a selective discretization, and propose an early warning system for very short-range heavy rainfall that employs the selective discretization. Finally, we draw conclusions and suggest future work in Chapter 5.

Chapter 2

Dimensional Reduction Techniques

When the number of features is unnecessarily large, the performance of machine learning algorithms may deteriorate due to the curse of dimensionality (Theodoridis and Koutroumbas, 2008). In addition, a large number of features in machine learning can significantly increase computational time and memory usage, and cause overfitting which leads to performance degradation on unseen data. There are two types of dimensionality reduction techniques to resolve this issue: feature selection and feature extraction (Li et al., 2017). Feature selection selects a subset of features, while feature extraction projects high-dimensional features to a lower dimensional space. Figure 2.1 shows the two types of dimensionality reduction techniques. When raw input data have little classification power, feature extraction tends to be preferred to feature selection (Abe, 2010; Li et al., 2017).

In this chapter, we introduce correlation-based feature selection, and principal component analysis which belongs to feature extraction. We then propose a scheme for precipitation type prediction as an illustrative example of an application of dimensionality reduction techniques.



(a) Feature selection (b) Feature extraction

Figure 2.1: Dimensionality reduction techniques

2.1 Correlation-based Feature Selection

The goal of feature selection is to obtain the subset of the input variables from which a classifier can be constructed which is small but discriminatory. Effective feature selection shortens training time and reduces overfitting. There are various criteria that can be used when selecting input variables (Kwak and Choi, 2002; Estévez et al., 2009; Hall, 1999; Peng et al., 2005). In this section, we describe the correlation-based feature selection (CFS) (Hall, 1999), which performed best in our preliminary experiments. The CFS uses some basic notions of information theory, which will be briefly covered. See Cover and Thomas (2006) for background information theory.

Let $(\mathbf{x}, y)$ be a tuple in which $\mathbf{x}$ contains values of independent random variables X_m $(1 \leq m \leq M)$, and y is the class label associated with the instance. That is another random variable Y, which can have one of the values: 1 to K, when there are K classes.

The entropy of a discrete random variable A is denoted by $\mathrm{H}(A)$. The random variable A has a set of possible values α and a probability mass function $p(a) = \Pr(A = a), a \in \alpha$. Then $\mathrm{H}(A)$ is defined by

$$\mathrm{H}(A) = -\sum_{a \in \alpha} p(a) \log p(a).$$

The entropy measures the uncertainty of a random variable, and the conditional entropy of two random variables A and B with a joint probability mass function $p(a, b)$ is defined as

$$\mathrm{H}(A|B) = -\sum_{a\in\alpha, b\in\beta} p(a,b) \log \frac{p(a,b)}{p(b)},$$

where β is a set of possible values for B, and $p(b)$ is a probability mass function of B. The conditional entropy $\mathrm{H}(A|B)$ quantifies the amount of randomness in A given the value of B.

The mutual information of two random variables A and B is denoted by $\mathrm{I}(A, B)$. The mutual information is calculated as

$$\begin{aligned}
\mathrm{I}(A,B) &= \sum_{a\in\alpha, b\in\beta} p(a,b) \log \frac{p(a,b)}{p(a)p(b)} \\
&= \sum_{a\in\alpha, b\in\beta} p(a,b) \log \frac{p(a|b)}{p(a)} \\
&= -\sum_{a\in\alpha, b\in\beta} p(a,b) \log p(a) + \sum_{a\in\alpha, b\in\beta} p(a,b) \log \frac{p(a,b)}{p(b)} \\
&= -\sum_{a\in\alpha} p(a) \log p(a) - \left(-\sum_{a\in\alpha, b\in\beta} p(a,b) \log \frac{p(a,b)}{p(b)} \right) \\
&= \mathrm{H}(A) - \mathrm{H}(A|B).
\end{aligned}$$

Thus, the mutual information quantifies the reduction in the uncertainty of one random variable through observing the other random variable.

The symmetric uncertainty of two random variables A and B is defined by

$$\mathrm{U}(A,B) = \frac{2 \cdot \mathrm{I}(A,B)}{\mathrm{H}(A) + \mathrm{H}(B)}.$$

The symmetric uncertainty normalizes mutual information so that its value lies between 0 and 1. The CFS uses the symmetric uncertainty to measure correlation between random variables.

The CFS is based on the hypothesis that a good feature subset contains features that are highly correlated with a class label, and largely uncorrelated with each other. Given a subset of features $S \subseteq \{X_1, X_2, \ldots, X_M\}$, merit of that subset is expressed as follows:

$$\text{Merit}_S = \frac{\sum_{X_i \in S} \text{U}(X_i, Y)}{\sqrt{\sum_{X_i \in S} \sum_{X_j \in S} \text{U}(X_i, X_j)}}.$$

The numerator measures the ability of S to predict the class label while the denominator measures the amount of information which is redundant between the selected features. To find the feature subset with the greatest merit, CFS uses the best-first search algorithm (Rich et al., 2009).

2.2 Principal Component Analysis

Principal component analysis (PCA) is a feature extraction technique that uses an orthogonal transformation to convert possibly correlated attributes into principal components (PCs) which are linearly uncorrelated attributes. The PCA provides an informative view of the data by introducing a new coordinate system and also provides a way to reduce the dimensionality of the data. For example, PCA is widely used to extract features from facial images (Cavalcanti et al., 2013), and can be used in wavelet denoising by discarding insignificant features from feature space (Yang and Ren, 2011).

Let $\mathbf{w}_{(p)} = (w_1, w_2, \ldots, w_M)$ be the p-th PC and $\mathbf{x}_n = (a_1, a_2, \ldots, a_M)$ be the attribute values of the n-th instance in a training set T, where M is the number of attributes associated with T and a_m is the value of the m-th attribute, for all $1 \leq m \leq M$. The first PC $\mathbf{w}_{(1)}$ is computed so that it has the largest variance:

$$\mathbf{w}_{(1)} = \underset{\|\mathbf{w}\|=1}{\text{argmax}} \sum_n (\mathbf{x}_n \cdot \mathbf{w})^2.$$

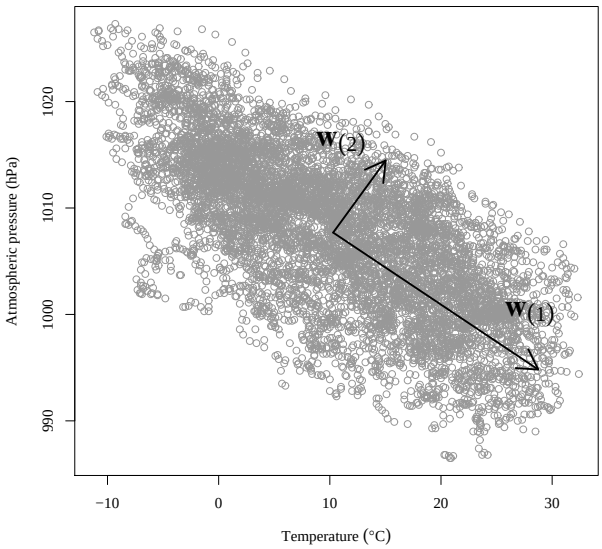


Figure 2.2: Illustration of PCA for two attributes: *Temperature* and *Atmospheric pressure*

The remaining PCs are in turn constructed so that they have the highest variance under the constraint that they should be orthogonal to the previous components. The number of PCs is less than or equal to M, and the dimensionality of the data is reduced by selecting the first s ($< M$) PCs without serious loss of information. After selecting s PCs, $\mathbf{x}_n$ is converted to $\hat{\mathbf{x}}_n = (\mathbf{w}_{(1)}\mathbf{x}_n, \mathbf{w}_{(2)}\mathbf{x}_n, \ldots, \mathbf{w}_{(s)}\mathbf{x}_n)$. The computation of the PCs can be carried out by the covariance method or the singular value decomposition (Jolliffe, 2002). With conventional implementations, the time complexity of PCA is $\mathcal{O}(M^2N)$, where M is the number of attributes and N is the number of instances. An illustrative example is shown in Figure 2.2.

The most common criterion for choosing s is using the cumulative percentage of total variation, i.e., s is determined as small as possible while the percentage of variation accounted for by the first s PCs is over the specified cutoff percentage. Although a sensible

cutoff lies in the range from 70 % to 90 %, it can be higher or lower depending on the properties of a data set (Jolliffe, 2002). In this study, we set the cutoff to 95 % through our preliminary experiments. Lowering the cutoff from 95 % to 90 % reduced s by 1.7 on average and decreased predictive performance. We standardized numerical attributes as standard practice (Murphy, 2012; Witten et al., 2016) and converted nominal attributes to binary numeric attributes before applying PCA. An attribute with q categorical values is converted to q binary (0 and 1) attributes, each of which indicates whether or not the value of the attribute falls into a certain category.

2.3 Case Study: Precipitation Type Forecast

Accurate prediction of precipitation types is important. In this section, we aim to improve the forecasting performance for three types (rain, snow, and sleet) of precipitation in South Korea during the winter season. A correlation-based feature selection method is used to select appropriate subsets of input variables, and multinomial logistic regression is used for classification of precipitation types. Comparative evaluations of various forecasting models are conducted using observational data from 2013 to 2015 for 22 major sites in South Korea.

2.3.1 Introduction

Surprisingly, precipitation type is not monitored in most meteorological stations and is often inaccessible (Liu et al., 2018). Most studies have focused on correctly classifying precipitation types when precipitation occurs, rather than forecasting precipitation types in advance. The most commonly used meteorological variables for classifying precipitation types are temperatures, and surface air temperature is usually the main predictor of precipitation types among various temperatures (Froidurot et al., 2014). Thus, there have been many studies that classify precipitation types using the threshold values of air temperature (Gao

et al., 2010; Kienzle, 2008; Lindström et al., 1997; Wigmosta et al., 1994; Yang et al., 1997) or the S-shaped curve that describes the relation between precipitation phase and the air temperature (Dai, 2008; Liu et al., 2018).

There have been studies that wet-bulb temperature is a better indicator than air temperature for discriminating precipitation types (Behrangi et al., 2018; Ding et al., 2014; Froidurot et al., 2014). Froidurot et al. (2014) reported that although the wet-bulb temperature is rarely measured and difficult to calculate for operational purposes, the use of the air temperature with relative humidity gave comparable performance to the wet-bulb temperature, and Behrangi et al. (2018) reported that the use of dew point temperature with the surface air temperature can be expected to provide good classification performance in precipitation types close to the wet-bulb temperature. On the contrary, there have also been reports that the wet-bulb temperature failed to provide an obvious advantage compared to the air temperature (Chen et al., 2014; Zhong et al., 2018).

Other than temperatures, there are meteorological variables that can affect the precipitation type. For example, relative humidity has an influence on precipitation types (Ding et al., 2014), wind speed helps to predict precipitation phase (Behrangi et al., 2018), surface pressure can affect the precipitation type at high elevations (Dai, 2008), the thicknesses of various pressure layers are used to differentiate precipitation types (Keeter and Cline, 1991), and vertical temperature lapse rate affects the precipitation phase (Sims and Liu, 2015).

A more complex model is needed when classifying precipitation types using multiple meteorological variables than when using only temperature thresholds. Logistic regression has been used to distinguish snowfall from rainfall using multiple meteorological variables. For example, Behrangi et al. (2018) first performed a principal component analysis (PCA) on input variables and classified precipitation types using logistic regression, Froidurot et al. (2014) used logistic regression on the data from 14 Swiss weather stations to determine precipitation types, and Jennings et al. (2018) produced Northern Hemisphere map of rain-

snow temperature thresholds by using a logistic regression on air temperature, relative humidity, and atmospheric pressure. Other than logistic regression, there have been studies that classified precipitation types using decision trees. For instance, Lee et al. (2014) used air temperature, relative humidity, and 1000-850 hPa thickness data to construct a decision tree that distinguishes rain, snow, and sleet in South Korea, and Reeves et al. (2016) produced a decision tree that categorizes six types of precipitation: rain, snow, sleet, freezing rain, ice pellets, and a freezing rain-ice pellet mix.

In this section, we introduce a novel method to improve the forecasting performance for precipitation types included in the short-range weather forecasts of European Centre for Medium-Range Weather Forecasts (ECMWF) and Regional Data Assimilation and Prediction System (RDAPS). We used various meteorological variables in the weather forecasts as input and performed feature selection techniques to select appropriate subsets of input variables to predict precipitation types. Multinomial logistic regression was used as classifiers, and comparative experiments were performed on various forecasting models. The experiments were conducted on 22 major sites in South Korea with 3-hourly lead times from 3 to 72 h. Empirical results showed that the proposed method improved prediction accuracy of ECMWF and RDAPS by more than 13 percentage points in predicting precipitation types in South Korea.

2.3.2 Forecast Model

Input Variables

Short-range weather forecasts from ECMWF and RDAPS are announced twice a day at 00:00 UTC and 12:00 UTC. Each forecast has 3-hourly resolution, i.e., it predicts weather elements after 3, 6, ..., and 72 h from the time of its publication. A total of 93 variables were used in this section, including various weather elements such as temperatures, wind speed, relative humidity, and precipitation type. Table 2.1 lists the input variables used in

this section. Both ECMWF and RDAPS forecasts use the same set of variables.

Table 2.1: Input variables for precipitation forecasts

No.	Variable	No.	Variable
01	Latitude of the target location (°)	02	Altitude of the target location (°)
03	Elevation of the target location (m)	04	Temperature at surface (K)
05	Temperature at 925 hPa (K)	06	Temperature at 850 hPa (K)
07	Temperature at 700 hPa (K)	08	Temperature at 500 hPa (K)
09	Relative humidity at surface (%)	10	Relative humidity at 925 hPa (%)
11	Relative humidity at 850 hPa (%)	12	Relative humidity at 700 hPa (%)
13	Relative humidity at 500 hPa (%)	14	Specific humidity at surface (kg/kg)
15	Specific humidity at 925 hPa (kg/kg)	16	Specific humidity at 850 hPa (kg/kg)
17	Specific humidity at 700 hPa (kg/kg)	18	Specific humidity at 500 hPa (kg/kg)
19	Dew point depression at surface (K)	20	Dew point depression at 925 hPa (K)
21	Dew point depression at 850 hPa (K)	22	Dew point depression at 700 hPa (K)
23	Dew point depression at 500 hPa (K)	24	East wind at surface (m/s)
25	East wind at 925 hPa (m/s)	26	East wind at 850 hPa (m/s)
27	East wind at 700 hPa (m/s)	28	East wind at 500 hPa (m/s)
29	South wind at surface (m/s)	30	South wind at 925 hPa (m/s)
31	South wind at 850 hPa (m/s)	32	South wind at 700 hPa (m/s)
33	South wind at 500 hPa (m/s)	34	North-east wind at surface (m/s)
35	North-east wind at 925 hPa (m/s)	36	North-east wind at 850 hPa (m/s)
37	North-east wind at 700 hPa (m/s)	38	North-east wind at 500 hPa (m/s)
39	North-west wind at surface (m/s)	40	North-west wind at 925 hPa (m/s)
41	North-west wind at 850 hPa (m/s)	42	North-west wind at 700 hPa (m/s)
43	North-west wind at 500 hPa (m/s)	44	Wind speed at surface (m/s)
45	Wind speed at 925 hPa (m/s)	46	Wind speed at 850 hPa (m/s)
47	Wind speed at 700 hPa (m/s)	48	Wind speed at 500 hPa (m/s)
49	Temperature max at surface (K)	50	Temperature min at surface (K)
51	Dew point temperature at surface (K)	52	Dew point temperature at 925 hPa (K)
53	Relative humidity at 300 hPa (%)	54	Dew point depression at 300 hPa (K)
55	Gust at surface (m/s)	56	Accumulated relative humidity at 925-500 hPa (%)
57	Accumulated relative humidity at 925-700 hPa (%)	58	Low cloud cover (%)
59	Total cloud cover (%)	60	Total column water vapour (kg/m^2)
61	Precipitable water at 500 hPa (kg/m^2)	62	Convective available potential energy (J/kg)
63	Precipitation (kg/m^2)	64	Snow (kg/m^2)
65	Equivalent potential temperature at 925 hPa (K)	66	Equivalent potential temperature at 850 hPa (K)

67	Equivalent potential temperature at 700 hPa (K)	68	Sky cover ({clear, scatter, broken, overcast})
69	Precipitation type ({rain, sleet, snow})	70	Depth of wet layer (DWL) at 1000-200 hPa (gpm)
71	Height of wet layer (HWL) at 1000-200 hPa (gpm)	72	Specific humidity of DWL at 1000-200 hPa (%)
73	Specific humidity of HWL at 1000-200 hPa (%)	74	Index for rainfall forecast at 1000-200 hPa
75	K-index	76	Lifted index
77	Showalter stability index	78	Lifted condensation level at 925 hPa (hPa)
79	Lifted condensation level at 850 hPa (hPa)	80	Lifted condensation level at 700 hPa (hPa)
81	Lapse rate at 850-500 hPa (°C/km)	82	Lapse rate at 850-700 hPa (°C/km)
83	Lapse rate at 925-850 hPa (°C/km)	84	Lapse rate at 950-850 hPa (°C/km)
85	Lapse rate at 950-925 hPa (°C/km)	86	Lapse rate at 1000-925 hPa (°C/km)
87	Potential vorticity at 850 hPa ($\mathrm{m^2\,s^{-1}\,K\,kg^{-1}}$)	88	Potential vorticity at 700 hPa ($\mathrm{m^2\,s^{-1}\,K\,kg^{-1}}$)
89	Potential vorticity at 500 hPa ($\mathrm{m^2\,s^{-1}\,K\,kg^{-1}}$)	90	Potential vorticity at 300 hPa ($\mathrm{m^2\,s^{-1}\,K\,kg^{-1}}$)
91	1000-850 hPa thickness (gpm)	92	1000-700 hPa thickness (gpm)
93	1000-500 hPa thickness (gpm)		

Classifier

We use multinomial logistic regression to predict precipitation types. Logistic regression is a classifier that is a type of regression analysis for binary classification. It assumes a linear relationship between the log odds of the dependent variable and the independent variables. Multinomial logistic regression is a generalization of logistic regression to multiclass problems, i.e., logistic regression can be used to classify rain and snow, while multinomial logistic regression can be used to classify rain, snow, and sleet. Let $(\mathbf{x}, y)$ be a tuple in which $\mathbf{x}$ contains values of independent random variables X_m $(1 \leq m \leq M)$, and y is the class label associated with the instance. That is another random variable Y, which can have one of the values: 1 to K, when there are K classes. The multinomial logistic regression chooses one class as a pivot and constructs $K - 1$ independent binary logistic regression models:

$$\ln \frac{\Pr(Y = k)}{\Pr(Y = K)} = \mathbf{b}_k \cdot \mathbf{x}, \tag{2.1}$$

where $\mathbf{b}_k$ is the set of regression coefficients associate with the class k for all $1 \leq k \leq K-1$, and class K is selected as the pivot. Then the probability that $\mathbf{x}_n$ belongs to the class k can be expressed as:

$$\Pr(Y = k) = \Pr(Y = K) e^{\mathbf{b}_k \cdot \mathbf{x}}, \tag{2.2}$$

for all $1 \leq k \leq K-1$. Since the sum of the probability that $\mathbf{x}_n$ belongs to each class is one, the probability that $\mathbf{x}_n$ belongs to the class K becomes

$$\begin{aligned} \Pr(Y = K) &= 1 - \sum_{k=1}^{K-1} \Pr(Y = k) \\ &= 1 - \sum_{k=1}^{K-1} \Pr(Y = K) e^{\mathbf{b}_k \cdot \mathbf{x}} \\ &= 1 - \Pr(Y = K) \sum_{k=1}^{K-1} e^{\mathbf{b}_k \cdot \mathbf{x}}. \end{aligned} \tag{2.3}$$

We can rewrite the above equation as follows:

$$\Pr(Y = K) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\mathbf{b}_k \cdot \mathbf{x}}}, \tag{2.4}$$

and Eq. (2.2) as follows:

$$\Pr(Y = k) = \frac{e^{\mathbf{b}_k \cdot \mathbf{x}}}{1 + \sum_{k=1}^{K-1} e^{\mathbf{b}_k \cdot \mathbf{x}}}. \tag{2.5}$$

Given observational data $\mathbf{x}$, multinomial logistic regression outputs a class label such that:

$$y = \underset{k}{\operatorname{argmax}} \Pr(Y = k). \tag{2.6}$$

The regression coefficients $\mathbf{b}_k$ are typically estimated by the maximum likelihood method (Hosmer et al., 2013). In this section, the ridge estimator (Cessie and Houwelingen, 1992) is used to prevent overfitting and unstable estimates. A feature with p categorical values is converted to p binary (0 or 1) features, each of which indicates whether or not the value of the feature falls into a certain category.

Target Areas

We have predicted precipitation types using short-range weather forecast data of 22 major sites in South Korea and verify them using cloud-to-ground lightning observational data which were recorded every 3 h. Figure 2.3 shows the location of each site. The major sites are representative cities and islands of South Korea, including the capital city Seoul. Detailed information such as the geographic location and wintertime precipitation types of each site is given in Table 2.2. Data from all the grid points are used as training data. Only wintertime precipitation data from January 1, 2013 to December 31, 2015 are used. In this section, wintertime means the period from December to February. Table 2.3 shows the occurrences of each precipitation type from 2013 to 2015 at the 22 major sites of South Korea. We can see that the majority of precipitation falls in the form of rain from March to November in South Korea.

Functionality

The forecast model for precipitation types aims to improve the prediction performance for precipitation types already included in the short-range weather forecasts of ECMWF and RDAPS. The precipitation type forecast is improved by using the weather elements forecast with the same forecast lead time, i.e., the precipitation type after t h is predicted using the weather elements forecast for that time. Therefore, the precipitation type forecast also has

Table 2.2: List of 22 major sites in South Korea

No.	Name	Lat. (°)	Lon. (°)	Alt. (m)	Occurrences in winter		
					Rain	Snow	Sleet
1	Chuncheon	37.9	127.7	77.7	77	134	13
2	Baengnyeongdo	38.0	124.6	144.9	64	195	29
3	Bukgangneung	37.8	128.9	78.9	75	156	21
4	Seoul	37.6	127.0	85.8	95	86	11
5	Incheon	37.5	126.6	71.4	87	89	14
6	Ulleungdo	37.5	130.9	222.8	173	508	84
7	Suwon	37.3	127.0	34.1	106	81	16
8	Seosan	36.8	126.5	28.9	107	140	34
9	Cheongju	36.6	127.4	57.2	114	123	17
10	Daejeon	36.4	127.4	68.9	141	120	36
11	Andong	36.6	128.7	140.1	96	66	4
12	Pohang	36.0	129.4	2.3	173	42	17
13	Daegu	35.9	128.7	49.0	64	10	1
14	Jeonju	35.8	127.1	61.4	161	78	25
15	Ulsan	35.6	129.3	34.6	170	35	17
16	Changwon	35.2	128.6	37.2	132	11	7
17	Gwangju	35.2	126.9	72.4	167	137	42
18	Busan	35.1	129.0	69.6	141	13	2
19	Mokpo	34.8	126.4	38.0	166	111	65
20	Yeosu	34.7	127.7	64.6	141	14	7
21	Heuksando	34.7	125.5	76.5	164	70	64
22	Jeju	33.5	126.5	20.5	371	53	84

The number of occurrences of each precipitation type was counted in 3-hourly winter observational data from 2013 to 2015.

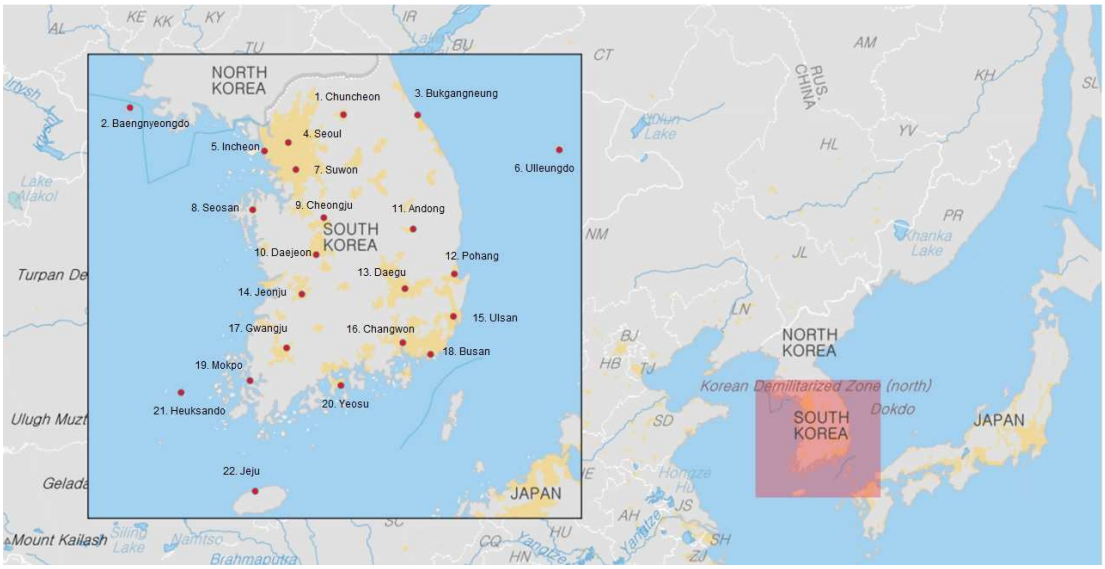


Figure 2.3: Locations of 22 sites in South Korea

Table 2.3: Monthly occurrences of each precipitation type at 22 sites

Month	Rain		Snow		Sleet	
Jan	867	(53 %)	626	(38 %)	153	(9 %)
Feb	1044	(57 %)	666	(36 %)	122	(7 %)
Mar	1181	(90 %)	90	(7 %)	39	(3 %)
Apr	2019	(99 %)	1	(0 %)	18	(1 %)
May	1289	(100 %)	0	(0 %)	0	(0 %)
Jun	1760	(100 %)	0	(0 %)	0	(0 %)
Jul	2782	(100 %)	0	(0 %)	0	(0 %)
Aug	2530	(100 %)	0	(0 %)	0	(0 %)
Sep	1683	(100 %)	0	(0 %)	0	(0 %)
Oct	1229	(100 %)	0	(0 %)	0	(0 %)
Nov	2233	(87 %)	220	(9 %)	112	(4 %)
Dec	1074	(45 %)	980	(41 %)	335	(14 %)

The number of occurrences of each precipitation type was counted in 3-hourly winter observational data from 2013 to 2015.

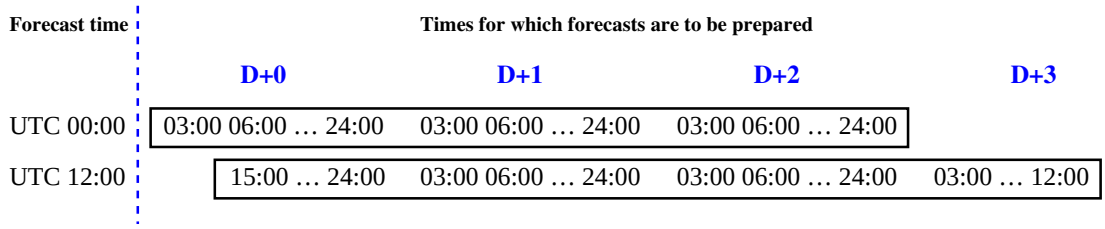


Figure 2.4: Forecast lead time for each forecast issuance time

3-hourly resolution, i.e., it predicts the type of precipitation after 3, 6, . . . , and 72 h from the time of its publication. Figure 2.4 shows the forecast lead time of the precipitation type prediction model.

Architecture

The architecture of the precipitation type forecast model is depicted in Figure 2.5. The forecast system is first trained on the meteorological database containing historical weather forecasts: data preprocessing methods and the classifier of the system are trained for precipitation type forecast. After training, the system takes input from ECMWF or RDAPS, and produces an output, the type of future precipitation. The data center updates the meteorological database with recent forecasts, and the forecast system can be retrained with the renewed database.

Performance Criteria

Accuracy, the number of correct predictions divided by the number of all predictions, is the main performance criterion for evaluating the overall performance of predicting precipitation types. In addition to accuracy, Heidke Skill Score (HSS) is used to evaluate the predictive performance for each precipitation type. Table 2.4 shows the contingency table

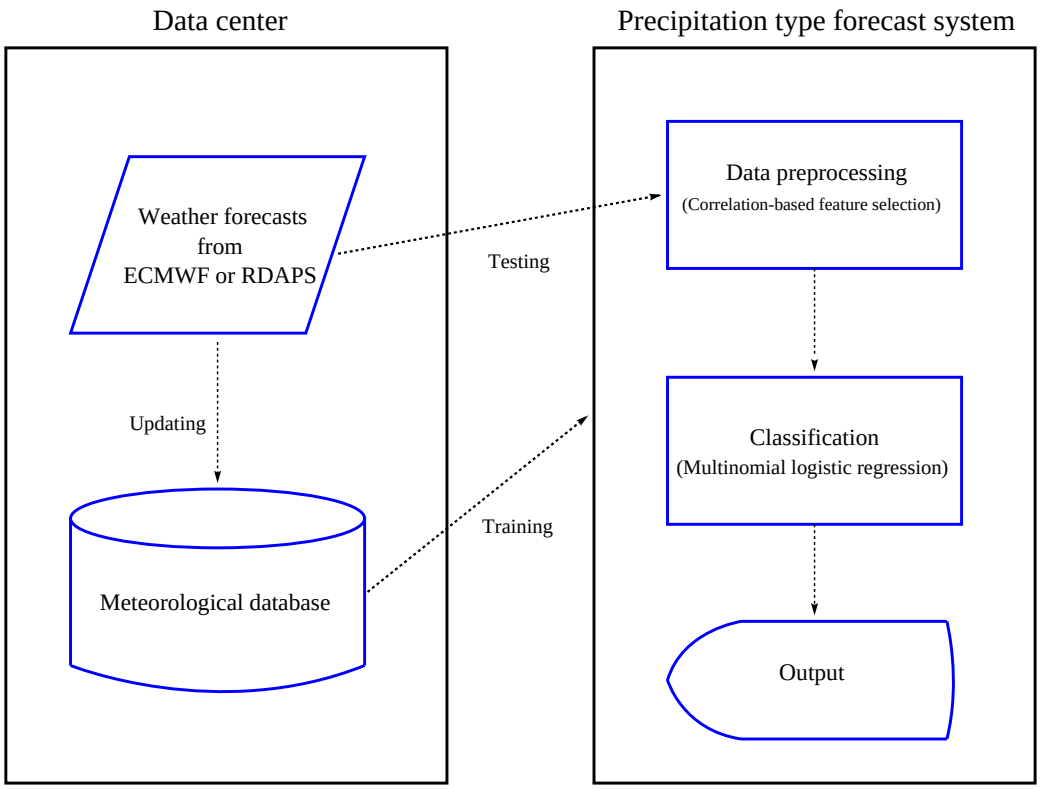


Figure 2.5: Architecture of the forecast model for precipitation types

Table 2.4: Confusion matrix for precipitation forecasts

Forecast	Observed		
	Yes	No	Total
Yes	a	b	$a+b$
No	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d=n$

of precipitation forecasts. Here, proportion correct (PC) for the event is:

$$\mathrm{PC} = \frac{a+d}{n}. \tag{2.7}$$

The HSS is an adjusted PC that is scaled by the portion of correct forecasts due to random chance in the absence of forecast skill. The probability that the predictions will hit by chance is:

$$E = \left(\frac{a+c}{n}\right)\left(\frac{a+b}{n}\right) + \left(\frac{b+d}{n}\right)\left(\frac{c+d}{n}\right), \tag{2.8}$$

which is the sum of the probabilities that a random forecast predicting *yes* is correct by chance and that a random forecast predicting *no* is correct by chance. Then, HSS is defined as:

$$\mathrm{HSS} = \frac{\mathrm{PC} - E}{1-E}. \tag{2.9}$$

Perfect forecast skill has an HSS value of 1 while no skill has a value of 0. The HSS was used by Behrangi et al. (2018) to compare various methods for determining precipitation phase. Please refer to Jolliffe and Stephenson (2003) and Wilks (2011) for general guidance on forecast verification including HSS.

2.3.3 Experiments

We conducted experiments to evaluate the performance of the proposed method, in which a feature selection is used to determine effective subsets of the input variables, and multinomial logistic regression was used for prediction.

Experiment Setup

We used the short-range weather forecast data of ECMWF and RDAPS for 22 major sites in South Korea from 2013 to 2015. A 3-fold cross-validation was performed to evaluate the performance of various forecast models for precipitation types. We set each fold to have annual data. To be specific, we used 2013 and 2014 data as training data to construct a model and 2015 data to evaluate the model in the first phase of the 3-fold cross-validation. We then used 2013 and 2015 for training and 2014 for testing in the second phase, and 2014 and 2015 for training and 2013 for testing in the third phase. The results of the three phases were averaged to produce a single estimation of performance for each forecasting model. The cross-validation experiments were conducted 3-hourly from 3 h to 72 h, depending on the forecast lead time, i.e., the precipitation type after k h was predicted using 93 input variables from the short-range weather forecast with the lead time of k h for all $k = 3, 6, \ldots, 72$. All experiments were performed separately using ECMWF data and RDAPS data to test the robustness of various forecast models and determine which dataset is better suited for precipitation type prediction.

As a baseline for predictive performances, the precipitation type variable (No. 69) included in the ECMWF and RDAPS short-range forecast was used. The results were also compared against those obtained using the improved Matsuo scheme (Lee et al., 2014). The scheme was generated by meteorologists to determine wintertime precipitation types in South Korea using air temperature, relative humidity, and 1000-850 hPa thickness. The models used as baselines do not need to be trained, so instead of the 3-fold cross-validation,

they were evaluated by the predictive performance over the entire period with the lead time of 3, 6, ..., and 72. We empirically show that multinomial logistic regression coupled with correlation-based feature selection (CFS) works well on precipitation type forecast through the comparative experiments of various forecast models.

Logistic regression coupled with PCA and decision tree models have been used to predict precipitation types with multiple input variables; thus, we also tested the performance of multinomial logistic regression preprocessed by PCA and a decision tree algorithm. The PCA is a feature extraction technique that uses an orthogonal transformation to convert possibly correlated features into linearly uncorrelated features. There have been studies that successfully applied PCA to logistic regression (Behrangi et al., 2018; Moon et al., 2019) in hydrological applications. As a decision tree algorithm, we used the C4.5, which is a decision tree learner that generates a decision tree using the concept of entropy in information theory. It builds a tree by recursively choosing the feature that best differentiates instances of the training set at each node of the tree. The C4.5 was selected for the top 10 algorithms in data mining (Wu et al., 2007). The improved Matsuo scheme used as a baseline can also be seen as a handcrafted decision tree. We implemented the improved Matsuo scheme in C language, and used the implementations of the WEKA (Waikato Environment for Knowledge Analysis) package (Hall et al., 2009) for the PCA, CFS, C4.5, and multinomial logistic regression.

Comparative Analysis

During non-winter seasons in South Korea, it is not difficult to predict precipitation types since precipitation occurs mostly in the form of rain. In winter, however, it is not easy to predict the precipitation type since various types of precipitation can occur. Table 2.5 shows the accuracy of precipitation type predictions included in the short-range forecasts of ECMWF and RDAPS. The average accuracies for all seasons were about 90 %, but they were around 70 % in winter. Therefore, it is necessary to improve the predictive accuracy

Table 2.5: Means and standard deviations of the accuracy of precipitation type predictions in ECMWF and RDAPS forecasts for all lead times

	All seasons	Winter only
ECMWF	0.8990 ± 0.0044	0.6854 ± 0.0164
RDAPS	0.9056 ± 0.0050	0.7077 ± 0.0112

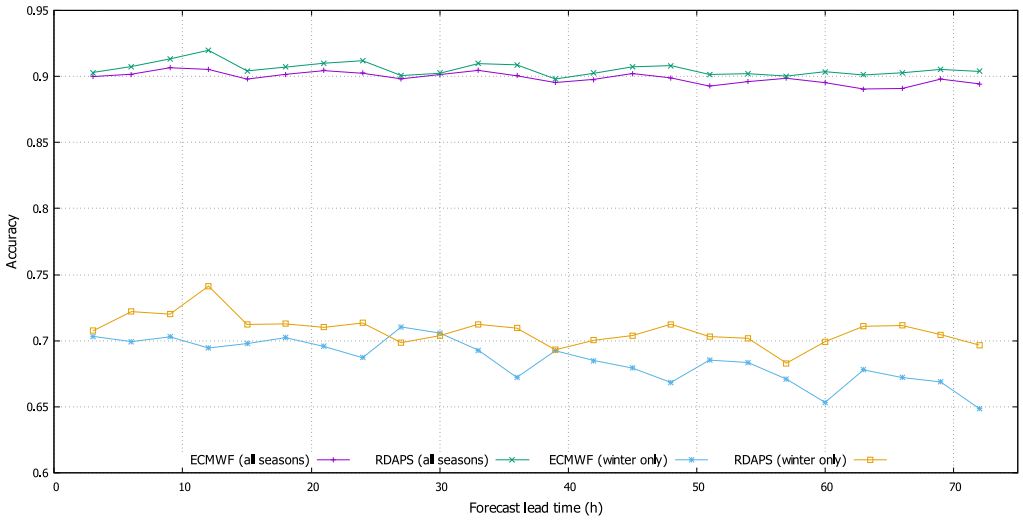


Figure 2.6: Accuracies of ECMWF and RDAPS for precipitation type predictions for different lead times

of precipitation types in winter. Figure 2.6 shows the accuracy of ECMWF and RDAPS predictions against lead time. We can see that as the forecast lead time increases, the accuracy tends to decrease, and RDAPS generally has higher accuracy than ECMWF in the precipitation type predictions. Wilcoxon signed-rank tests were performed on the accuracy pairs of ECMWF and RDAPS for all forecast lead times to see if their accuracies have the same distribution. The tests indicated that RDAPS was more accurate than ECMWF both in all seasons and winter ($p < 0.001$).

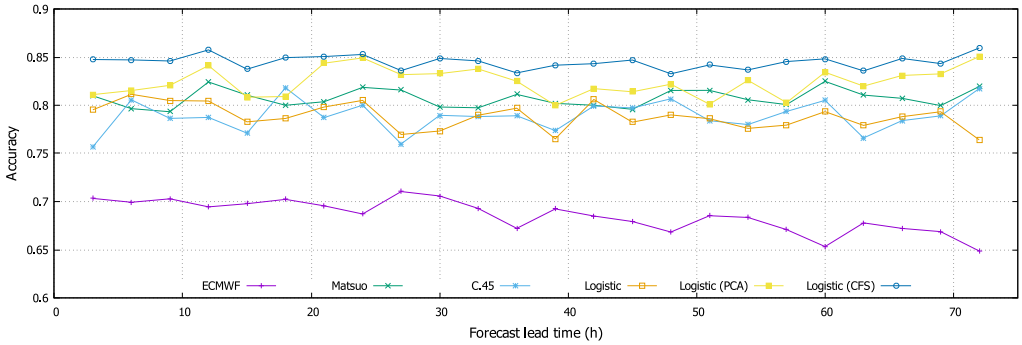
Experiments were conducted using ECMWF short-range forecast data to predict wintertime precipitation types in South Korea. Figure 2.7 shows the performance of six models

Table 2.6: Performance comparison of the proposed method and the other methods in ECMWF dataset

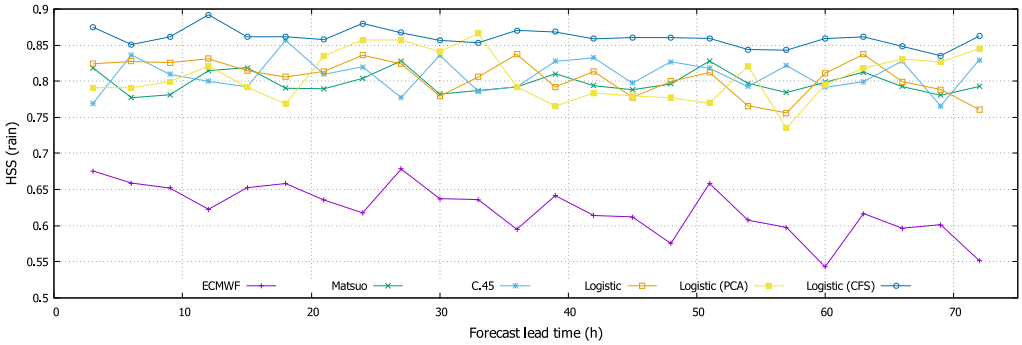
	Accuracy		HSS (rain)		HSS (snow)		HSS (sleet)	
	Average	p-value	Average	p-value	Average	p-value	Average	p-value
Logistic (CFS)	**0.8449**		**0.8603**		**0.7346**		0.1705	
ECMWF	0.6854	< 0.001	0.6224	< 0.001	0.5189	< 0.001	0.0674	< 0.001
Matsuo	0.8073	< 0.001	0.7983	< 0.001	0.7127	< 0.001	0.0759	< 0.001
C4.5	0.7890	< 0.001	0.8090	< 0.001	0.6341	< 0.001	**0.1716**	0.8887
Logistic	0.7884	< 0.001	0.8058	< 0.001	0.6460	< 0.001	0.1705	0.8887
Logistic (PCA)	0.8241	< 0.001	0.8067	< 0.001	0.7027	< 0.001	0.1380	0.0016

For each measure, the highest average value is shown in bold type.

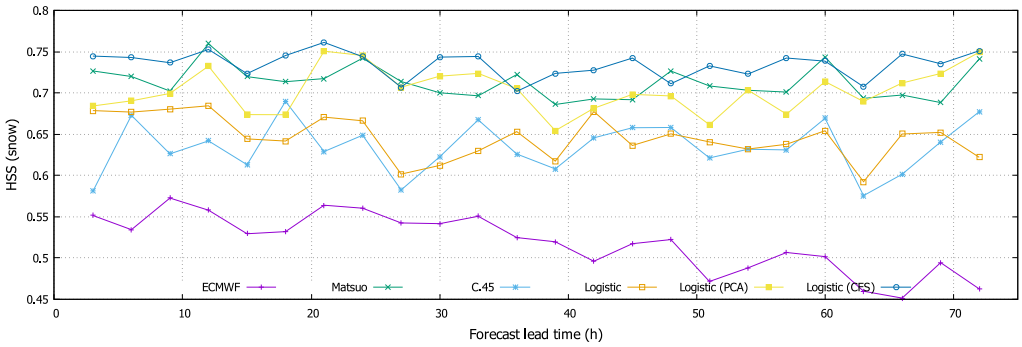
predicting precipitation types using ECMWF input dataset. Since the input has the precipitation type variable (No. 69) of ECMWF, a sensible model is expected to have at least as much accuracy as ECMWF. We can see that all models outperformed ECMWF except that the performance of the improved Matsuo scheme was sometimes lower than that of ECMWF in the sleet forecast. The proposed method, which is the multinomial logistic regression combined with CFS, has the highest performance and the multinomial logistic regression with PCA has the second highest performance for all measures except the HSS for sleet. Table 2.6 compares the proposed method with the other methods with respect to the accuracy and the HSS of each precipitation type. The table gives the average value and the result of the Wilcoxon signed-rank test for each performance criterion. The null hypothesis of the Wilcoxon signed-rank test is that the medians of the performance measure are the same between the proposed method and the compared method. The lower the p-value is, the more significant the difference between the performances of the two methods is. The table shows that the proposed method outperforms the other methods with a significant difference in the accuracy and the HSS for rain and snow. In the case of sleet, the performance of C4.5 was the best, and the proposed method was the second best. However, the p-value shows no significant difference in performance between the two.



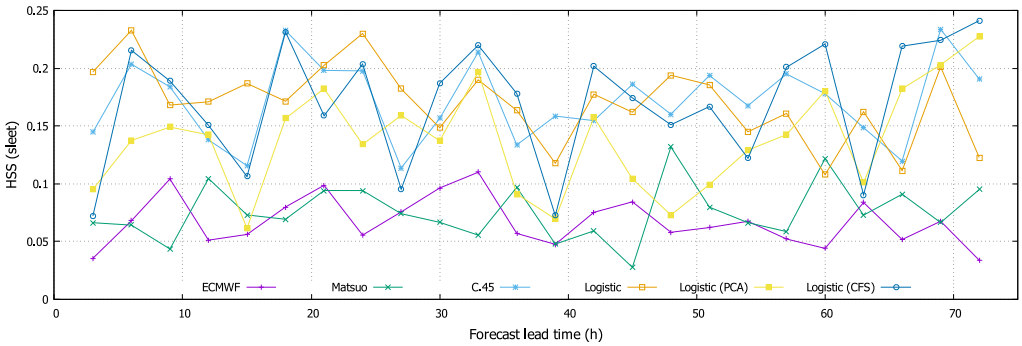
(a) Accuracy



(b) HSS for rain



(c) HSS for snow



(d) HSS for sleet

Figure 2.7: Comparison of wintertime precipitation type predictions using ECMWF data

'Logistic' denotes the multinomial logistic regression and the technique in parentheses represents the data preprocessing method.

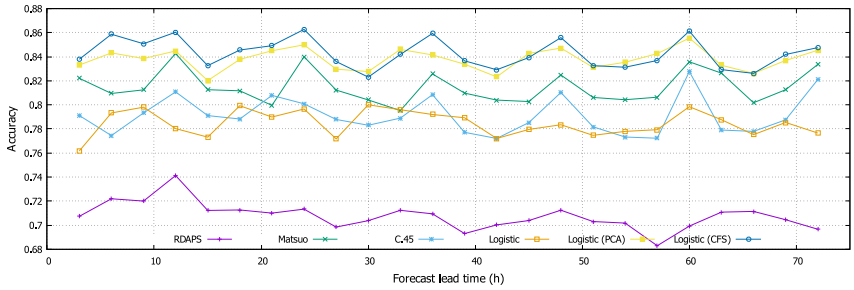
Table 2.7: Performance comparison of the proposed method and the other methods in RDAPS dataset

	Accuracy		HSS (rain)		HSS (snow)		HSS (sleet)	
	Average	p-value	Average	p-value	Average	p-value	Average	p-value
Logistic (CFS)	**0.8428**		**0.8554**		**0.7303**		**0.1934**	
RDAPS	0.7077	< 0.001	0.6197	< 0.001	0.5547	< 0.001	0.1086	< 0.001
Matsuo	0.8149	< 0.001	0.7997	< 0.001	0.7173	0.0434	0.0636	< 0.001
C4.5	0.7913	< 0.001	0.8147	< 0.001	0.6386	< 0.001	0.1589	0.0244
Logistic	0.7847	< 0.001	0.8019	< 0.001	0.6340	< 0.001	0.1702	0.0466
Logistic (PCA)	0.8379	0.0041	0.8512	0.0989	0.7294	0.6965	0.1825	0.1236

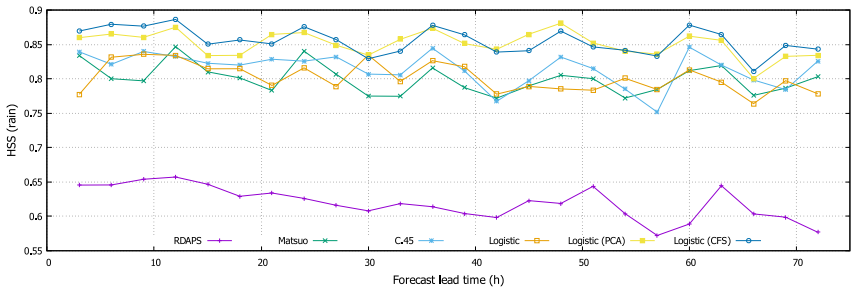
For each measure, the highest average value is shown in bold type.

Figure 2.8 shows the performance of six models predicting precipitation types using RDAPS input dataset. Similar to ECMWF, the performances of most prediction models were better than that of RDAPS. In addition, we can see that the overall performances of multinomial logistic regression combined with CFS or PCA were superior to the other models. However, the performance of the improved Matsuo scheme was worse than that of RDAPS in the sleet forecast. Table 2.7 shows that the proposed method has the highest average score in all performance measures compared to the other methods. The results of the Wilcoxon signed-rank test showed that the proposed method outperforms the the other methods with the significance level at 0.05 except the multinomial logistic regression with PCA. In the HSS for each precipitation type, the proposed method has higher mean score than the multinomial logistic regression with PCA but did not show a statistically significant difference on RDAPS dataset. However, the accuracy of the proposed method was significantly better than that of the multinomial logistic regression with PCA.

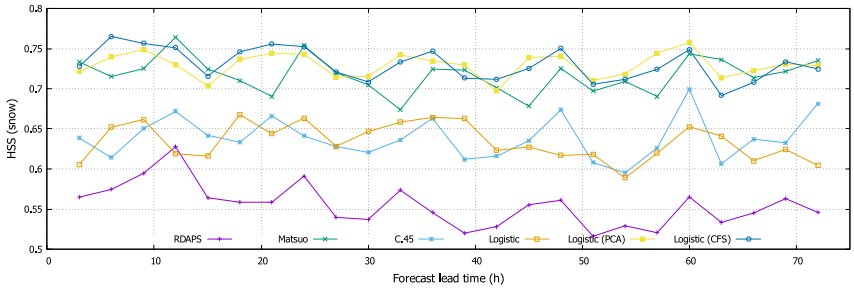
The proposed method showed the best performance in both ECMWF and RDAPS datasets. We compared the performance of the proposed method in ECMWF and RDAPS to see which one is better for predicting precipitation types. The Wilcoxon signed-rank tests were conducted with the null hypothesis that there is no difference in the performance of



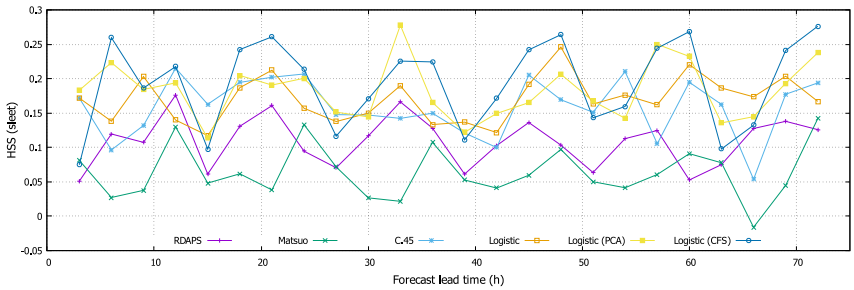
(a) Accuracy



(b) HSS for rain



(c) HSS for snow



(d) HSS for sleet

Figure 2.8: Comparison of wintertime precipitation type predictions using RDAPS data

'Logistic' denotes the multinomial logistic regression and the technique in parentheses represents the data preprocessing method.

the proposed method depending on the input dataset. Table 2.8 compares the performance of the proposed method according to the input dataset. When ECMWF was used as input, the average of each performance criterion except the HSS for sleet was higher, but there was no statistically significant difference. In the case of the sleet forecast, RDAPS showed better performance than ECMWF when used as input of the proposed method. The table also shows that the HSSs of the sleet are much lower than those of the other precipitation types. The performance degradation that occurs only in a certain class is common in imbalanced data where the number of instances of one class is much smaller than the others (He and Garcia, 2009; Su et al., 2006; Sun et al., 2007). To achieve high overall predictive accuracy, most machine learning algorithms exhibit a high predictive performance for prevalent class instances, but poor performance for minority class instances (Cardie and Howe, 1997; Chawla et al., 2002). The portion of sleet instances is only 9.13 % of the total precipitation instances, which seems to have lowered the predictive performance for sleet. To be specific, there were 3,510 observations of sleet during the experiment period. However, the proposed method produced 1,520 predictions for sleet in ECMWF dataset and 1,691 in RDAPS dataset. It is necessary to resolve this problem when the sleet forecast is far more important than the other precipitation types. In this case, however, the predictive performance for the other precipitation types will become worse generally.

Predictive Performance for Each Site

Table 2.9 gives the performance of the proposed method on ECMWF dataset for each site, and Table 2.10 gives the performance on RDAPS dataset for each site. Since our experiments did not distinguish between the sites, the more frequent precipitation the site had, the more the prediction performance of the site was reflected in the results. There seems to be no significant difference in the predictive performance of each site according to datasets. For

Table 2.8: Performance of the proposed method on each dataset

	Dataset		p-value
	ECMWF	RDAPS	
Accuracy	0.8449 ± 0.0070	0.8428 ± 0.0122	0.2757
HSS (rain)	0.8603 ± 0.0121	0.8554 ± 0.0190	0.1188
HSS (snow)	0.7346 ± 0.0157	0.7303 ± 0.0200	0.2113
HSS (sleet)	0.1705 ± 0.0525	$\mathbf{0.1934 \pm 0.0634}$	0.0096

Each row shows the mean, standard deviation, and the p-value of the Wilcoxon signed-rank test for all lead times.

example, the accuracies for Mokpo and Heuksando were lower than 80 %, but the accuracies for Daegu, Changwon, Busan, and Yeosu were higher than 90 % on both datasets. The accuracies for the other sites were between 80 % and 90 %. Although PCs for sleet were generally higher than those for snow, HSSs for sleet were much lower than those for snow in most sites. The HSS measures the proportion of improvements over random chance. In the case of sleet, it is possible to achieve a high PC with random forecasts that predict that sleet will not occur with high probability since sleet rarely occurred in many sites. Therefore, it is considered that the sleet forecasts of the proposed method are better than random forecasts only in some sites. In addition, large differences in the forecasting performances for sleet by the sites are due to the fact that the number of sleet occurrences is relatively small compared to the other precipitation types. For example, sleet occurred only once in Daegu on our dataset, and thus HSS for sleet can vary greatly depending on the success of the prediction for the event. In the case of rain and snow, however, the proposed method has good forecast skill since PCs and HSSs had high values in all the sites.

Table 2.9: Predictive performance of precipitation types for 22 major sites in South Korea using the ECMWF dataset

No.	Name	Accuracy	Rain		Snow		Sleet	
			PC	HSS	PC	HSS	PC	HSS
1	Chuncheon	0.8631	0.8969	0.7735	0.8915	0.7708	0.9377	0.0006
2	Baengnyeongdo	0.8364	0.9369	0.8297	0.8450	0.6333	0.8909	0.0011
3	Bukgangneung	0.8175	0.8879	0.7331	0.8407	0.6510	0.9064	0.0014
4	Seoul	0.8807	0.9171	0.8343	0.9053	0.8107	0.9390	0.0005
5	Incheon	0.8862	0.9422	0.8842	0.9101	0.8201	0.9202	0.0368
6	Ulleungdo	0.8473	0.9266	0.8086	0.8858	0.7381	0.8822	0.0268
7	Suwon	0.8495	0.9071	0.8136	0.8822	0.7605	0.9097	0.0001
8	Seosan	0.8080	0.9113	0.8177	0.8352	0.6716	0.8695	0.0724
9	Cheongju	0.8505	0.9201	0.8391	0.8717	0.7440	0.9092	0.0007
10	Daejeon	0.8241	0.9309	0.8619	0.8442	0.6865	0.8731	0.0109
11	Andong	0.8880	0.9108	0.8128	0.9015	0.7893	0.9637	0.0006
12	Pohang	0.8113	0.8749	0.6765	0.8644	0.5460	0.8833	0.1950
13	Daegu	0.9505	0.9646	0.7858	0.9505	0.6687	0.9858	0.2447
14	Jeonju	0.8528	0.9366	0.8662	0.8698	0.7101	0.8993	0.0015
15	Ulsan	0.8533	0.9310	0.7945	0.8947	0.5735	0.8809	0.0733
16	Changwon	0.9135	0.9476	0.7070	0.9338	0.5016	0.9455	0.0009
17	Gwangju	0.8423	0.9623	0.9245	0.8601	0.7120	0.8621	0.0539
18	Busan	0.9421	0.9746	0.8365	0.9492	0.5454	0.9604	0.2208
19	Mokpo	0.7630	0.9523	0.9046	0.7834	0.5074	0.7904	0.2274
20	Yeosu	0.9452	0.9831	0.9204	0.9562	0.7260	0.9512	0.2451
21	Heuksando	0.7896	0.9780	0.9552	0.8070	0.4814	0.7942	0.3355
22	Jeju	0.8311	0.9165	0.7797	0.8930	0.4973	0.8526	0.3639

Table 2.10: Predictive performance of precipitation types for 22 major sites in South Korea using the RDAPS dataset

No.	Name	Accuracy	Rain		Snow		Sleet	
			PC	HSS	PC	HSS	PC	HSS
1	Chuncheon	0.8746	0.9154	0.8155	0.8992	0.7884	0.9346	0.0072
2	Baengnyeongdo	0.8278	0.9350	0.8224	0.8339	0.6069	0.8866	0.0087
3	Bukgangneung	0.8155	0.8806	0.7130	0.8414	0.6508	0.9091	0.0151
4	Seoul	0.8725	0.9126	0.8252	0.8980	0.7958	0.9344	0.0011
5	Incheon	0.8817	0.9422	0.8840	0.8963	0.7931	0.9248	0.1008
6	Ulleungdo	0.8431	0.9200	0.7940	0.8840	0.7356	0.8822	0.0212
7	Suwon	0.8538	0.9209	0.8411	0.8788	0.7529	0.9080	0.0020
8	Seosan	0.8150	0.9252	0.8454	0.8327	0.6669	0.8720	0.0966
9	Cheongju	0.8587	0.9270	0.8531	0.8730	0.7467	0.9174	0.0065
10	Daejeon	0.8093	0.9103	0.8205	0.8353	0.6686	0.8731	0.0178
11	Andong	0.8703	0.8942	0.7778	0.8838	0.7512	0.9627	0.0006
12	Pohang	0.8148	0.8728	0.6849	0.8623	0.5823	0.8945	0.2066
13	Daegu	0.9410	0.9552	0.7711	0.9552	0.7414	0.9717	0.0004
14	Jeonju	0.8509	0.9287	0.8496	0.8620	0.6930	0.9111	0.0991
15	Ulsan	0.8613	0.9237	0.7817	0.9041	0.6471	0.8947	0.1169
16	Changwon	0.9177	0.9487	0.7169	0.9359	0.5369	0.9509	0.0308
17	Gwangju	0.8304	0.9539	0.9076	0.8497	0.6891	0.8571	0.0494
18	Busan	0.9502	0.9736	0.8378	0.9604	0.6862	0.9665	0.2188
19	Mokpo	0.7645	0.9558	0.9116	0.7869	0.5097	0.7864	0.2440
20	Yeosu	0.9422	0.9801	0.9086	0.9522	0.7292	0.9522	0.0943
21	Heuksando	0.7745	0.9710	0.9413	0.7936	0.4589	0.7843	0.2968
22	Jeju	0.8314	0.9149	0.7704	0.9063	0.3879	0.8417	0.4179

Analysis of Feature Selection

We trained a different model for each forecast lead time using the proposed method and investigated which input variables were selected by the CFS in each model. Table 2.11 lists the input variables selected by the CFS for more than half of the prediction models. Regardless of the input datasets, snow (No. 64) and the thickness at 1000-700 hPa (No. 92) variables were selected by all prediction models, and features related to humidity, temperature, and wind are selected with a high probability. Among the features indicating the target location, only the latitude was selected by more than half of the prediction models. When using ECMWF dataset, there were more models that did not use precipitation type (No. 69) variable than models that used the variable. On RDAPS dataset, however, most models used the precipitation type variable. On average, the proposed method used 16 input variables in ECMWF dataset, and 18 variables in RDAPS dataset.

2.3.4 Discussions

Accurate prediction of precipitation types is important for the study of water resources assessments, land hydrological processes, and road traffic safety. However, it is not easy to predict the precipitation type in winter due to its chaotic characteristics. To improve forecasting performance for precipitation types included in the short-range weather forecasts of ECMWF and RDAPS, we presented a novel method to classify rain, snow, and sleet using machine learning techniques.

Most of the existing methods have classified precipitation types using only a small number of input variables, mainly based on temperatures. We used 93 meteorological variables of the short-range weather forecasts as input and selected a subset of relevant variables through CFS. Various features such as snow, thickness, humidity, temperature, wind speed, and latitude of the target location were selected for prediction of precipitation types. We believe

Table 2.11: List of selected input variables by CFS

Variable name	ECMWF	RDAPS	Avg.
Snow (kg/m^2)	1.0000	1.0000	1.0000
Thickness of geopotential height at 1000-700 hPa (gpm)	1.0000	1.0000	1.0000
Thickness of geopotential height at 1000-850 hPa (gpm)	0.9722	1.0000	0.9861
Specific humidity at surface (kg/kg)	0.9167	0.9306	0.9236
Thickness of geopotential height at 1000-500 hPa (gpm)	0.9306	0.9028	0.9167
Equivalent potential temperature at 925 hPa (K)	0.8889	0.9444	0.9167
Temperature at 850 hPa (K)	0.8889	0.8889	0.8889
North-east wind at surface (m/s)	0.7639	0.7778	0.7708
Latitude of the target location (°)	0.6806	0.7639	0.7222
South wind at 500 hPa (m/s)	0.5833	0.8056	0.6944
Precipitation type ({rain, sleet, snow})	0.4583	0.8750	0.6667
Relative humidity at 500 hPa (%)	0.6944	0.4861	0.5903
Dew point depression at 500 hPa (K)	0.3889	0.6528	0.5208
Specific humidity at 850 hPa (kg/kg)	0.4583	0.5694	0.5139

Selected ratios in each dataset together with averages are shown.
The variables selected by more than half of the prediction models on average were listed among 93 input variables.

that this is the first study that applied feature selection to precipitation type forecasts.

After feature selection, multinomial logistic regression were used to classify wintertime precipitation types in South Korea. The comparative analysis of various forecasting models was conducted to show that the proposed method works well on precipitation type forecast. The proposed method had the highest performance in the experiments and improved prediction accuracies of ECMWF and RDAPS by more than 15 percentage points and 13 percentage points, respectively. In addition, the proposed method showed higher predictive performance than the improved Matsuo scheme which is specialized in precipitation type forecasts for South Korea.

Sleet had the lowest HSS among the three types of precipitation. It seems to be due to a relatively small number of instances compared to the other precipitation types. Our future work aims to improve the sleet forecasting performance through undersampling (Liu et al., 2009), oversampling (Chawla et al., 2002) or boosting (Sun et al., 2007) techniques while minimizing the performance degradation for the other precipitation types.

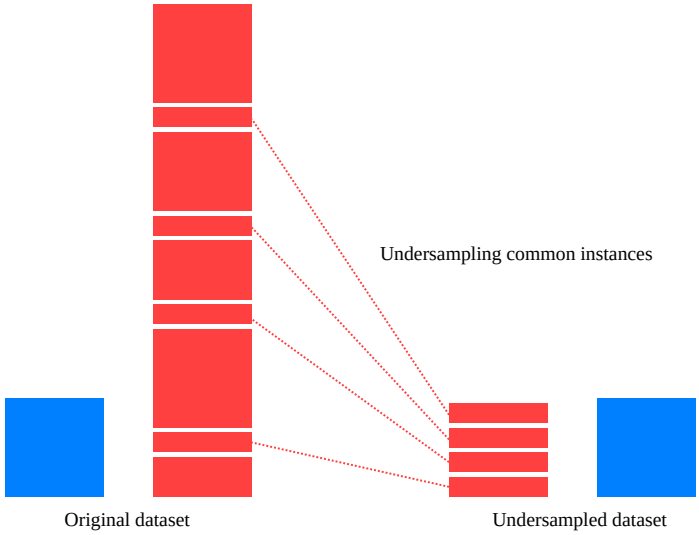
Chapter 3

Sampling Techniques

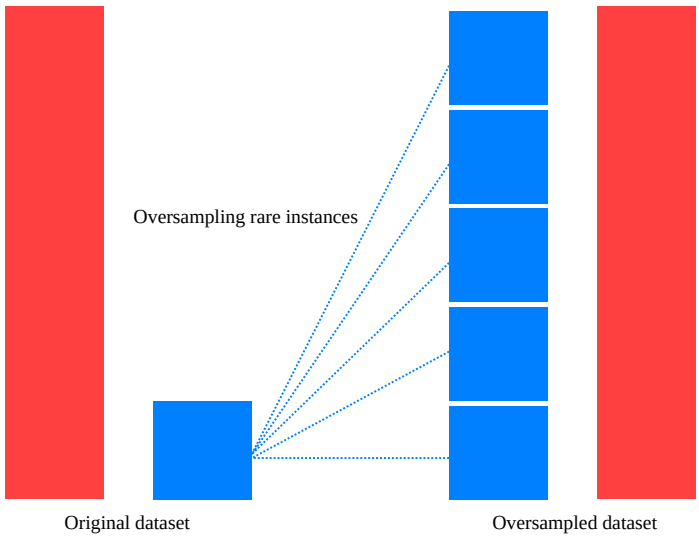
Sometimes the goal of machine learning is to predict unusual events. In this case, the number of rare instances in the training data are far less than that of common instances. Most machine learning algorithms sacrifice performance on rare instances to overall performance (Cardie and Howe, 1997; Chawla et al., 2002), which is referred to as class imbalance problem. To balance the class distribution, undersampling reduces the number of common instances, and oversampling increases the number of rare instances. Figure 3.1 compares undersampling and oversampling. In this chapter, we describe the two sampling techniques, and propose a lightning forecast model, which balances the class distribution to predict the rare atmospheric phenomenon.

3.1 Undersampling

Undersampling is one of the popular methods in dealing with the imbalanced data (Liu et al., 2009). Since undersampling reduces the size of the training data, it is preferred over oversampling when dealing with big data. In this section, we describe random undersampling, which is easy to implement and effective in class imbalance learning. Japkowicz (2000) reported



(a) Undersampling



(b) Oversampling

Figure 3.1: Undersampling and oversampling

that more sophisticated sampling methods than random undersampling were unnecessary.

Let $T = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ be the training set with N instances in which $\mathbf{x}_n$ contains values of independent variables, and y_n is the dependent variable, which can have one of two values, where '-1' signifies a *minority* class, and '1' signifies a *majority* class. In addition, we define subsets $T_{min} \subset T$, where T_{min} is the set of all minority instances and T_{maj} is the set of all majority instances. Random undersampling randomly selects a set of instances $S \subset T_{maj}$ and constructs a new training set T_{und} so that $T_{und} = T_{min} \cup S$. There is no way to know in advance how many majority instances should be removed by undersampling for the optimal result. Therefore, we vary the rates of undersampling, as in the study of Estabrooks et al. (2004) and Dubey et al. (2014), to find the optimal undersampling ratio.

3.2 Oversampling

While undersampling removes common instances from the original dataset, oversampling appends rare instances to the original dataset. In this section, we introduce random oversampling and synthetic oversampling. The random oversampling duplicates rare instances, and the synthetic oversampling generates new instances.

Random oversampling randomly selects a rare instance with replacement, and append it to the training dataset. This process is repeated until the number of rare instances grows to the desired number. Similar to random undersampling, we do not know in advance how many rare instances should be duplicated for the optimal result. Therefore, the number of rare instances to be duplicated should be determined by trial and error. Since random oversampling simply appends replicated data to the original dataset, it is known to be vulnerable to overfitting (He and Garcia, 2009).

Synthetic oversampling generates new rare instances, and appends them to the original dataset. As an example of synthetic oversampling, we describe the synthetic minority over-

sampling technique (SMOTE) due to Chawla et al. (2002). For simplicity, we assume that all independent variables of the instances are continuous. For each rare instance $t_i \in T_{min}$, SMOTE first finds k nearest neighbours that belong to T_{min}. The SMOTE then randomly selects one of the k nearest neighbours, and synthesize a new instance t_{new} as follows:

$$t_{new} = t_i + \delta \cdot (\hat{t_i} - t_i),$$

where $\hat{t_i}$ is the selected neighbor of t_i and $\delta \in [0, 1]$ is a random number. This process is repeated until the number of rare instances reaches a predefined number. The number of iteration and k are hyperparameters whose values are set before oversampling. The SMOTE showed promising results on various datasets (Chawla et al., 2002), however it has drawbacks such as over generalization and the overlapping problem between classes (He and Garcia, 2009).

3.3 Case Study: Lightning Forecast

Accurate prediction of lightning activities is important to minimize risks to life and property. In this section, we construct a prediction model to forecast lightning activities around the Korean Peninsula with undersampling and support vector machines (SVMs). Short-range weather forecasts from European Centre for Medium-range Weather Forecasts (ECMWF) for various meteorological variables were used as input and forecast lead times were 3-hourly from 9 to 30 h. Since the occurrence of lightning activities is not common, imbalanced learning problem should be handled when we apply machine learning techniques to lightning forecast. We use undersampling technique to resolve the skewed data distribution and employed the SVMs to predict lightning activities. Since it is very difficult to predict exactly when and where lightning activity occurs, we extend the spatial and temporal scales of lightning predictions to improve the predictive performance of the proposed method.

3.3.1 Introduction

Traditionally, lightning was predicted using parameters that are highly correlated with lightning activities. For example, Price and Rind (1992) used convective cloud top height (CLDHT) in the lightning parameterization, Allen and Pickering (2002) parameterized lightning flash rates in terms of CLDHT, convective precipitation, and upward convective mass flux, Bright et al. (2005) utilized convective available potential energy (CAPE) to devise a physical-based parameter for lightning prediction, and Yair et al. (2010) introduced the lightning potential index parameter which is the kinetic energy of the updraft in the developing thundercloud.

Data assimilation has been widely employed in numerical weather prediction (NWP) to forecast lightning activities (Fierro et al., 2014; Lynn et al., 2015; Giannaros et al., 2016). Given current weather conditions, the NWP uses mathematical models to simulate the atmosphere and forecasts the future state of the weather. This method, however, is not appropriate for regional forecast because of spin-up problems and its coarse spatial and temporal resolution (Mecklenburg et al., 2000). An alternative approach is needed to complement the NWP on a smaller spatial and temporal scale.

The performance of a machine learning algorithm is usually measured by accuracy. Therefore, it is common practice to evaluate classifiers by the rate of correct classification and regression functions by the mean squared error. When data set is highly imbalanced, however, the overall prediction accuracy may be misleading (Chawla et al., 2002; He and Garcia, 2009; Liu et al., 2009; Su et al., 2006; Sun et al., 2007). For example, only 1.5 % instances of our dataset needed to be classified as a lightning instance. A learning algorithm may decide to classify all instances as non-lightning instances so that it can achieve 98.5 % accuracy, which is not desirable. Thus, lightning forecast needs a proper performance criterion other than the accuracy.

In machine learning, undersampling (He and Garcia, 2009; Liu et al., 2009) is one of

the techniques used to increase the predictive performance of learning algorithms in highly imbalanced data. Imbalanced data means that the number of majority instances is much larger than that of minority instances. Undersampling alleviates the imbalanced data problem by removing samples from the majority class, which can improve the training speed of learning algorithms and the predictive performance for the minority class.

This section investigates the possibility of employing machine learning techniques in forecasting lightning activities. Meteorological variables made available in the short-range weather forecast from European Centre for Medium-range Weather Forecasts (ECMWF) provide the input to our scheme, in which an undersampling is used to make training dataset more balanced, and support vector machines (SVMs) (Cortes and Vapnik, 1995) are used for classification. Experiments were conducted on the Korean Peninsula and its surrounding areas with 3-hourly lead times from 9 to 30 h.

3.3.2 Forecast Model

Input Variables

The ECMWF short-range weather forecasts are announced twice a day at 00:00 UTC and 12:00 UTC. Each forecast predicts weather variables at 3-hour intervals, from 9 to 30 h ahead. The 112 weather variables, which are listed in Table 3.1, include temperatures, wind speed, relative humidity, CAPE, K-index, and Showalter stability index.

Table 3.1: Input variables for lightning forecasts

No.	Variable	No.	Variable
01	Month	02	Temperature at surface (K)
03	Temperature at 925 hPa (K)	04	Temperature at 850 hPa (K)
05	Temperature at 700 hPa (K)	06	Temperature at 500 hPa (K)
07	Relative humidity at surface (%)	08	Relative humidity at 925 hPa (%)
09	Relative humidity at 850 hPa (%)	10	Relative humidity at 700 hPa (%)
11	Relative humidity at 500 hPa (%)	12	Specific humidity at surface (kg/kg)
13	Specific humidity at 925 hPa (kg/kg)	14	Specific humidity at 850 hPa (kg/kg)
15	Specific humidity at 700 hPa (kg/kg)	16	Specific humidity at 500 hPa (kg/kg)
17	Dew point depression at surface (K)	18	Dew point depression at 925 hPa (K)
19	Dew point depression at 850 hPa (K)	20	Dew point depression at 700 hPa (K)
21	Dew point depression at 500 hPa (K)	22	East wind at surface (m/s)
23	East wind at 925 hPa (m/s)	24	East wind at 850 hPa (m/s)
25	East wind at 700 hPa (m/s)	26	East wind at 500 hPa (m/s)
27	South wind at surface (m/s)	28	South wind at 925 hPa (m/s)
29	South wind at 850 hPa (m/s)	30	South wind at 700 hPa (m/s)
31	South wind at 500 hPa (m/s)	32	North-east wind at surface (m/s)
33	North-east wind at 925 hPa (m/s)	34	North-east wind at 850 hPa (m/s)
35	North-east wind at 700 hPa (m/s)	36	North-east wind at 500 hPa (m/s)
37	North-west wind at surface (m/s)	38	North-west wind at 925 hPa (m/s)
39	North-west wind at 850 hPa (m/s)	40	North-west wind at 700 hPa (m/s)
41	North-west wind at 500 hPa (m/s)	42	Wind speed at surface (m/s)
43	Wind speed at 925 hPa (m/s)	44	Wind speed at 850 hPa (m/s)
45	Wind speed at 700 hPa (m/s)	46	Wind speed at 500 hPa (m/s)
47	Temperature max at surface (K)	48	Temperature min at surface (K)
49	Dew point temperature at surface (K)	50	Dew point temperature at 925 hPa (K)
51	Relative humidity at 300 hPa (%)	52	Dew point depression at 300 hPa (K)
53	Gust at surface (m/s)	54	Accumulated relative humidity at 925-500 hPa (%)
55	Accumulated relative humidity at 925-700 hPa (%)	56	Low cloud cover (%)
57	Total cloud cover (%)	58	Total column water vapour (kg/m^2)
59	Precipitable water at 500 hPa (kg/m^2)	60	Convective available potential energy (J/kg)
61	Precipitation (kg/m^2)	62	Snow (kg/m^2)
63	Equivalent potential temperature at 925 hPa (K)	64	Equivalent potential temperature at 850 hPa (K)
65	Equivalent potential temperature at 700 hPa (K)	66	Sky cover ({clear, scatter, broken, overcast})
67	Precipitation type ({rain, sleet, snow})	68	Depth of wet layer (DWL) at 1000-200 hPa (gpm)
69	Height of wet layer (HWL) at 1000-200 hPa (gpm)	70	Specific humidity of DWL at 1000-200 hPa (%)

71	Specific humidity of HWL at 1000-200 hPa (%)	72	Index for rainfall forecast at 1000-200 hPa
73	K-index	74	Lifted index at 500 hPa
75	Parcel lifted index at 500 hPa	76	Showalter stability index at 850-500 hPa
77	Lifted condensation level at 925 hPa (hPa)	78	Lifted condensation level at 850 hPa (hPa)
79	Lifted condensation level at 700 hPa (hPa)	80	Lapse rate at 850-500 hPa (°C/km)
81	Lapse rate at 850-700 hPa (°C/km)	82	Lapse rate at 925-850 hPa (°C/km)
83	Lapse rate at 950-850 hPa (°C/km)	84	Lapse rate at 950-925 hPa (°C/km)
85	Lapse rate at 1000-925 hPa (°C/km)	86	Potential vorticity at 850 hPa ($m^2\,s^{-1}\,K\,kg^{-1}$)
87	Potential vorticity at 700 hPa ($m^2\,s^{-1}\,K\,kg^{-1}$)	88	Potential vorticity at 500 hPa ($m^2\,s^{-1}\,K\,kg^{-1}$)
89	Potential vorticity at 300 hPa ($m^2\,s^{-1}\,K\,kg^{-1}$)	90	Total Totals index
91	1000-700 hPa thickness (gpm)	92	Maximum temperature at 850 hPa (K)
93	Minimum temperature at 850 hPa (K)	94	Dew point temperature at 850 hPa (K)
95	Dew point temperature at 700 hPa (K)	96	Dew point temperature at 500 hPa (K)
97	Showalter stability index at 925-500 hPa	98	Showalter stability index at 925-700 hPa
99	Lifted index at 925 hPa	100	Averaged lifted index
101	Convective condensation level (m)	102	Temperature at convective condensation level (K)
103	Convective temperature (K)	104	Storm relative helicity (m^2/s^2)
105	Lifted condensation level (m)	106	Temperature at lifted condensation level (K)
107	Level of free convection (m)	108	Temperature at level of free convection (K)
109	Equilibrium level (m)	110	Temperature at equilibrium level (K)
111	Convective inhibition (J/kg)	112	Total precipitable water (kg/m^2)
113	Freezing level (m)		

Classifier

We use support vector machines (SVMs) to predict lightning. The SVMs are supervised learning techniques that can be used for classification and regression analysis. The original SVMs were first introduced in 1963 (Vapnik and Lerner, 1963), but they were considered as alternatives to artificial neural networks in the 1990s since nonlinear classification became possible through the kernel trick (Boser et al., 1992). While conventional classifiers minimize error rates during the training process, the SVMs construct a set of hyperplanes so that the distance from it to the nearest training data point is maximized. We will briefly introduce SVMs in the following paragraphs. For details on SVMs, refer to Burges (1998) and Ivanciuc (2007).

Suppose that we have a training set $T = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ and want to find the maximum-

margin hyperplane that divides the set of points $\mathbf{x}_n$ for which $y_n = -1$ from that for which $y_n = 1$. If the training set is linearly separable, a hyperplane that separates training examples according to their class labels can be written as $\mathbf{w} \cdot \mathbf{x} + w_0 = 0$, where $\mathbf{w}$ is the normal vector to the hyperplane, and $|w_0| \,/\, \|\mathbf{w}\|$ is the distance from the hyperplane to the origin. With a normalized dataset, we can find a hyperplane satisfying the following two constraints:

$$\mathbf{w} \cdot \mathbf{x}_n + w_0 \geq 1, \text{ for } y_n = 1 \tag{3.1}$$

and

$$\mathbf{w} \cdot \mathbf{x}_n + w_0 \leq -1, \text{ for } y_n = -1. \tag{3.2}$$

Each constraint is a hyperplane that separates the instances of the corresponding class label, and the distance between the two hyperplanes is $2 \,/\, \|\mathbf{w}\|$. Figure 3.2 demonstrates the maximum-margin hyperplanes in a two-dimensional case. Now, we can define the following optimization problem:

$$\text{minimize } \|\mathbf{w}\|^2 \text{ subject to } y_n(\mathbf{w} \cdot \mathbf{x}_n + w_0) - 1 \geq 0, \text{ for all } 1 \leq n \leq N \tag{3.3}$$

to find the maximum-margin hyperplane that has the largest separation between the two classes.

When the training set is not linearly separable, i.e., if all constraints in Eq. (3.3) cannot be satisfied, we introduce the hinge loss function to penalize the misclassified instances:

$$\max(0, 1 - y_n(\mathbf{w} \cdot \mathbf{x}_n + w_0)). \tag{3.4}$$

This function penalizes the instances on the wrong side of the margin with the value of the

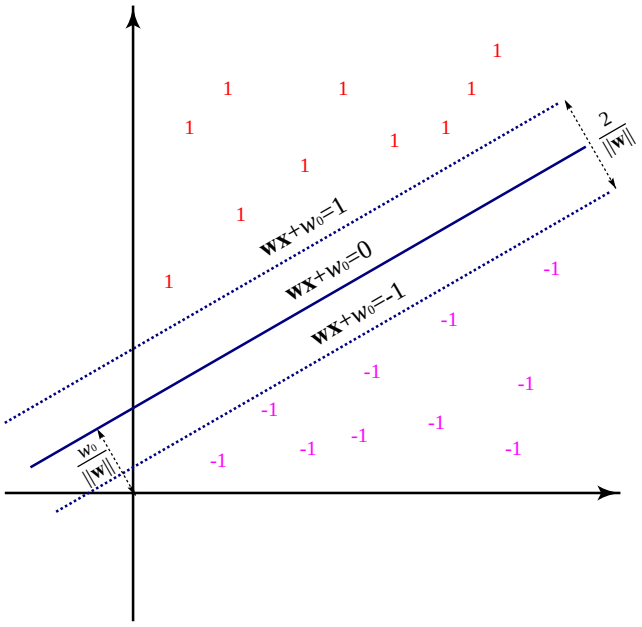


Figure 3.2: Maximum-margin hyperplanes separating instances according to their class labels

distance from the margin. Then we define the following optimization problem to maximize the margin while reducing misclassified instances:

$$\text{minimize} \quad \left[\frac{1}{N}\sum_{n=1}^{N}\max(0, 1 - y_n(\mathbf{w}\cdot\mathbf{x}_n + w_0))\right] + \lambda\|\mathbf{w}\|^2, \tag{3.5}$$

where λ is a parameter that controls the tradeoff between increasing the margin and reducing misclassified instances. In practice, we can solve this problem in $O(N^2)$ time using the sequential minimal optimization proposed by Platt (1999).

Target Areas

Figure 3.3 shows the map of areas covered by this section. Starting at (31°N, 123°E), there are 2,400 grid points with 60 grids in the North direction and 40 grids in the East direction. The interval between the grids is 0.25°, and thus the end of the grid points is (46°N, 133°E).

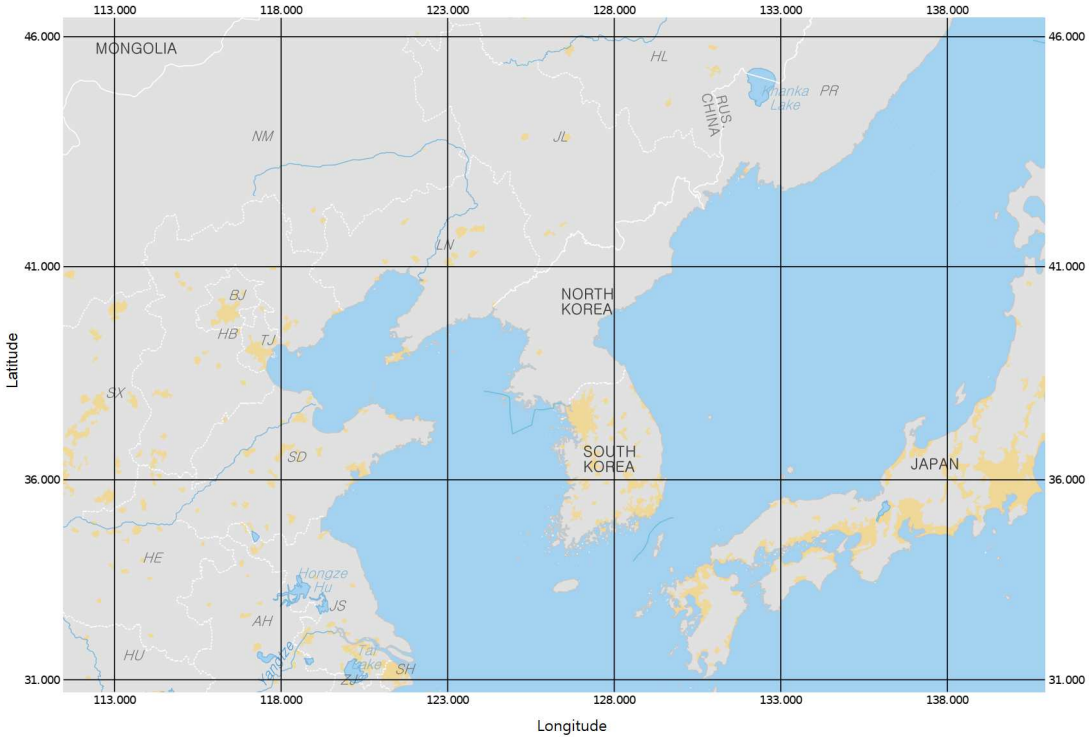


Figure 3.3: Map around the Korean Peninsula

Figure 3.4 shows the frequency of lightning activities from 2015 to 2016 at all the grid points: lightning occurred intensively from April to September around the Korean Peninsula. Thus, the experiment period was set from April to September in 2015 and 2016, and grids that experienced less than 10 lightning during this period were excluded from the experiment to avoid severe data imbalance. As a result, only 1,161 grid points were used in our experiments.

Functionality

The lightning forecast system predicts lightning at 3-hour intervals, from 9 to 30 h ahead. The lightning activities during each interval are predicted from the weather variables forecast for that interval. Figure 3.5 shows the forecast lead time of the lightning forecast model.

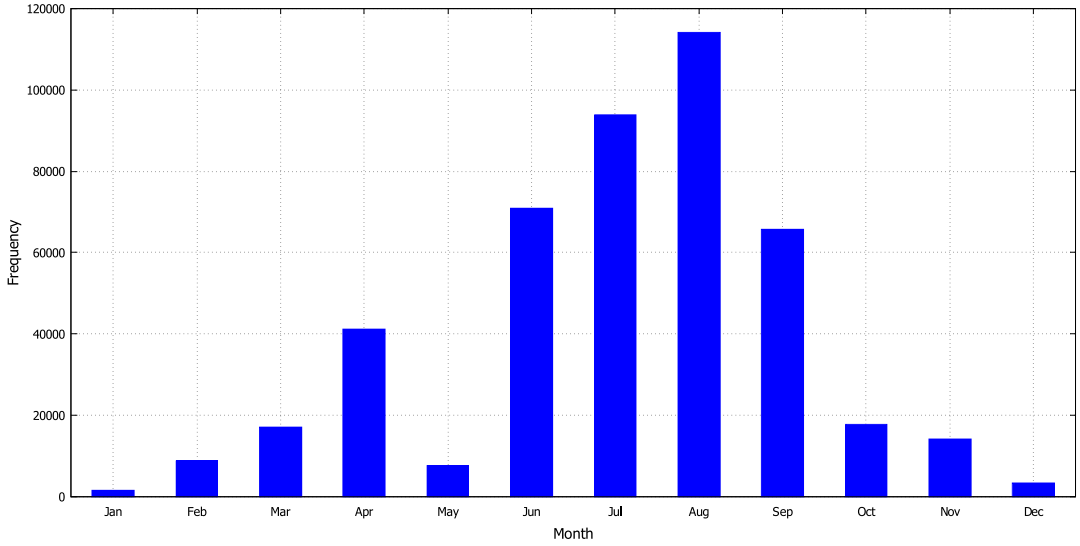


Figure 3.4: Monthly frequency of lightning activities on our dataset

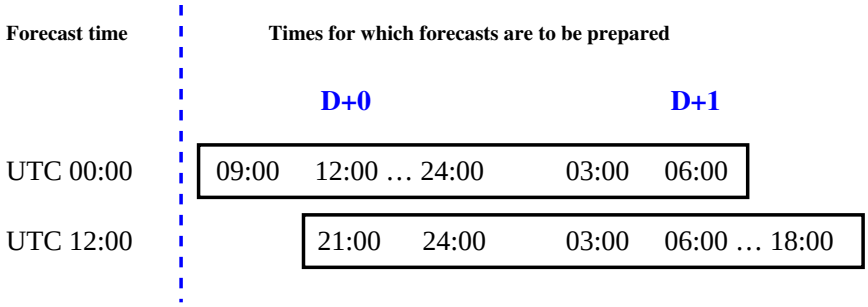


Figure 3.5: Forecast lead time for each forecast issuance time

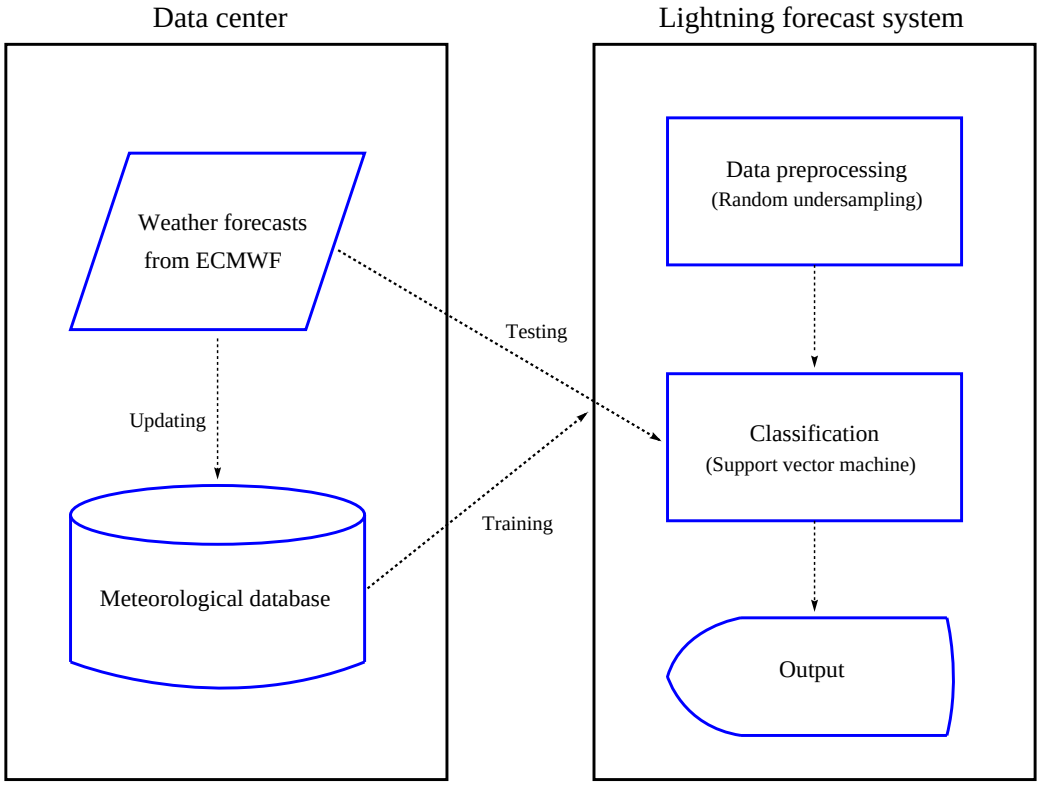


Figure 3.6: Architecture of the lightning forecast model

Architecture

The architecture of the lightning forecast model is depicted in Figure 3.6. The forecast system is first trained on the meteorological database containing historical weather forecasts. After training, the system takes input from ECMWF, and produces an output, whether or not lightning will occur within a particular location and time interval. The data center updates the meteorological database with recent forecasts, and the forecast system can be retrained with the renewed database.

Table 3.2: Confusion matrix for lightning forecasts

Forecast	Observed	
	Yes	No
Yes	True positive (TP)	False positive (FP)
No	False negative (FN)	True negative (TN)

Performance Criteria

A confusion matrix is typically used to visualize the prediction results for a binary classification, as shown in Table 3.2. From the confusion matrix, we compute probability of detection (POD), false alarm ratio (FAR), and equitable threat score (ETS) as follows:

$$\mathrm{POD} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad \mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}},$$

and

$$\mathrm{ETS} = \frac{\mathrm{TP} - \alpha}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} - \alpha},$$

where α denotes the expected number of correct forecasts by chance:

$$\alpha = \frac{(\mathrm{TP} + \mathrm{FP}) \cdot (\mathrm{TP} + \mathrm{FN})}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}.$$

Among the three measures, ETS is used as the main performance criterion since it is a balanced measure that takes account into both FP and FN. The ETS is more reliable than critical success index (also known as threat score), which does not consider random forecasts. Therefore, ETS is a commonly used metric to evaluate the performance of lightning forecast (Dafis et al., 2018; Giannaros et al., 2017; Lynn et al., 2015). Refer to Jolliffe and Stephenson (2003) and Wilks (2011) for general guidance on forecast verification.

3.3.3 Experiments

We conducted experiments to evaluate the performance of the proposed method, in which an undersampling is used to make training dataset more balanced, and SVMs are used to predict lightning.

Experiment Setup

We used the short-range weather-forecast data around the Korean Peninsula from 2015 to 2016. The total amount of the data was approximately 8.3 GB, and experiments were conducted on the E3-1225 processor with a clock rate of 3.2 GHz. A two-fold cross-validation was performed to evaluate the different methods of forecasting lightning activities. The 2015 data was used to train a model, which was then evaluated on the 2016 data. This procedure was repeated using the 2016 data for training and the 2015 data for evaluation. The results from these procedures were averaged to produce a single estimation of performance for each forecasting model. The cross-validation was repeated for each 3-hour interval, from 9 to 30 h after the forecast time.

We tested two representative classifiers to determine which one is better suited for lightning prediction. The classifiers used in our experiments were SVMs and random forests (Breiman, 2001). Random forests are an ensemble learning technique that constructs multiple decision trees on various subsamples of the training data and takes a majority vote to classify an instance. It can be seen as a bagging algorithm (Breiman, 1996) for decision trees that uses only a random subset of features for splitting each node. The bagging helps learning algorithms to improve the predictive accuracy and to avoid overfitting. We used the Waikato Environment for Knowledge Analysis (WEKA) package (version 3.8.1) due to Hall et al. (2009) to implement these classifiers, in which we used the default settings of the package: random forests made up 100 decision trees, and SVMs normalized input data and used a polynomial kernel.

Table 3.3: Results on 9 h lightning forecasts for different classifiers and undersampling ratios

Classifier	Ratio[a]	TP	FP	FN	TN	POD	FAR	ETS
SVMs	1:1	10,350	165,789	2,473	671,240	**0.8071**	0.9412	0.0437
	1:5	5,496	36,068	7,327	800,961	0.4286	0.8696	**0.1007**
	1:10	2,117	10,171	10,706	826,858	0.1651	**0.8277**	0.0847
RFs[b]	1:1	8,482	103,736	4,341	733,293	0.6615	0.9244	0.0591
	1:5	3,640	21,485	9,183	815,544	0.2839	0.8555	0.0957
	1:10	1,808	9,087	11,015	827,942	0.1410	0.8341	0.0756

[a] # of lightning instances : # of non-lightning ones.
[b] Random forests.
The best values for each performance criterion are shown in bold type.

Comparative Analysis

We conducted experiments on 9 h lightning forecasts to find an appropriate undersampling ratio. The 9 h lightning forecast predicts whether or not lightning activity occurs during the 9–12 h interval after the forecast time. Table 3.3 shows the results of the SVMs and random forests for three undersampling ratios. The undersampling ratio indicates how many non-lightning instances are included in the training data against the number of lightning instances. At the 1:1 ratio, in which the number of lightning instances is the same as that of non-lightning instances, the classifiers predicted lightning with a high probability, resulting in the highest POD and FAR. At the 1:10 ratio, however, the POD and FAR were the lowest as the classifiers were less likely to predict lightning. At the 1:5 ratio, where the number of non-lightning instances is five times greater than that of lightning instances, the ETS was the highest because the POD and FAR were balanced. Therefore, all subsequent experiments performed the undersampling at the 1:5 ratio. Section 3.3.3 contains experimental results for various undersampling ratios, in which the 1:5 ratio appears to be optimal.

Table 3.4 compares the predictive performance of SVMs and random forests and gives the POD, FAR, and ETS for each classifier and forecast lead time. Random forests had a

Table 3.4: Comparative performance for different classifiers and forecast lead times

Lead time	POD		FAR		ETS	
	SVMs	RFs	SVMs	RFs	SVMs	RFs
09 h	**0.4507**	0.2827	0.8696	**0.8555**	**0.1007**	0.0957
12 h	**0.2922**	0.1942	0.8979	**0.8936**	**0.0717**	0.0650
15 h	**0.4206**	0.2691	0.8817	**0.8687**	**0.0910**	0.0875
18 h	**0.4451**	0.2966	0.8575	**0.8399**	**0.1081**	0.1050
21 h	**0.3988**	0.2348	0.8773	**0.8708**	**0.0918**	0.0810
24 h	**0.2594**	0.1826	0.9052	**0.8939**	**0.0646**	0.0632
27 h	**0.3748**	0.2260	0.8856	**0.8828**	**0.0855**	0.0745
30 h	**0.3871**	0.2501	0.8708	**0.8540**	**0.0944**	0.0907
Average	**0.3786**	0.2420	0.8807	**0.8699**	**0.0885**	0.0828
Standard deviation	0.0692	0.0406	0.0155	0.0195	0.0144	0.0147

The best values for each performance criterion are shown in bold type.

lower value of FAR than SVMs, but SVMs had higher values of POD and ETS than random forests. Specifically, random forests produced about 1 % fewer false alarms than SVMs, but SVMs successfully detected about 13 % more lightning activities than random forests, and thus achieved higher ETS values than random forests. However, even SVMs overall had ETS values less than 0.1, which indicates that it is very difficult to predict lightning activities in the current temporal and spatial scale.

Performance over the Land and the Sea

The lightning parameterization scheme implemented by Dafis et al. (2018) showed better performance over the land than over the sea. We investigated whether or not the proposed method in this section also has different performances over the land and the sea. Figure 3.7 shows the map of areas that are mainly classified by administrative region. In this map, the grids on zero belong to the sea and the rest belong to the land. Among the 1,161 grids used in this section, 764 grids belong to the sea and the remaining 397 grids belong to the land.

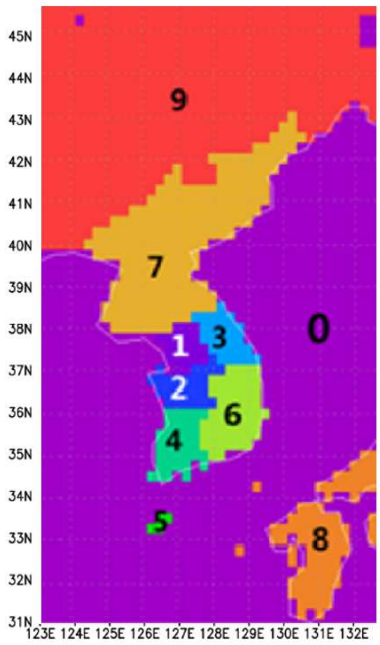
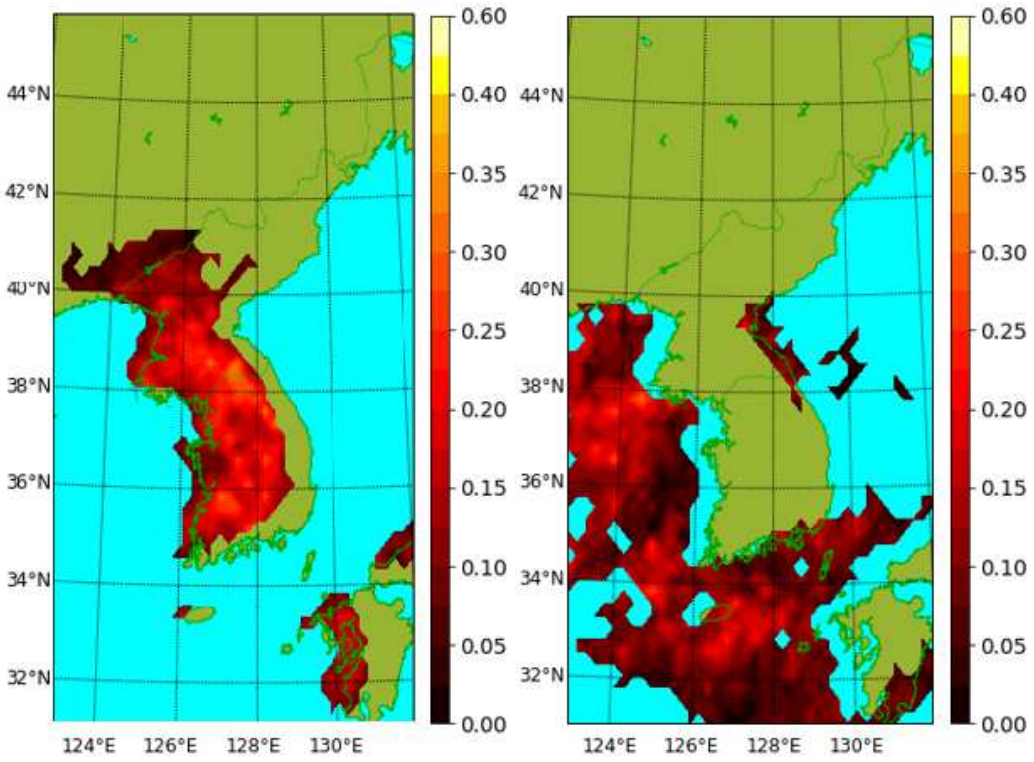


Figure 3.7: Map of areas primarily categorized by administrative district

Table 3.5 compares the predictive performance of the proposed method over the land and the sea. In the cross-validation process, the land models used only the data from the land grids, and the sea models used only the data from the sea grids. The sea models were slightly better than the land models in terms of FAR, but the land models were generally better than the sea models in terms of POD and ETS. To be specific, the sea models produced about 3 % less false alarms than the land models, but the land models successfully detected about 9 % more lightning, and thus achieved 0.02 higher ETS than the sea models. Figure 3.8 compares the ETS values for land models and sea ones. Overall, the proposed method showed higher predictive performance over the land than over the sea.



(a) ETS over the land.

(b) ETS over the sea.

Figure 3.8: Performance comparison of land models and sea ones

Table 3.5: Comparison of predictive performance by target area

Lead time	POD		FAR		ETS	
	Land	Sea	Land	Sea	Land	Sea
09 h	**0.5450**	0.4275	0.8604	**0.8165**	**0.1106**	0.0903
12 h	0.3025	**0.3112**	0.8965	**0.8509**	**0.0734**	0.0702
15 h	**0.5143**	0.3630	0.8602	**0.8405**	**0.1114**	0.0765
18 h	**0.5141**	0.3996	0.8347	**0.8144**	**0.1263**	0.0912
21 h	**0.4868**	0.3739	0.8732	**0.8377**	**0.0974**	0.0775
24 h	0.2425	**0.291**	0.9089	**0.8568**	0.0609	**0.0668**
27 h	**0.4583**	0.3380	0.8660	**0.8472**	**0.1037**	0.0727
30 h	**0.4701**	0.3387	0.8455	**0.8379**	**0.1149**	0.0771
Average	**0.4417**	0.3554	0.8682	**0.8377**	**0.0998**	0.0778
Standard deviation	0.1091	0.0451	0.0246	0.0153	0.0221	0.0088

The best values for each performance criterion are shown in bold type.

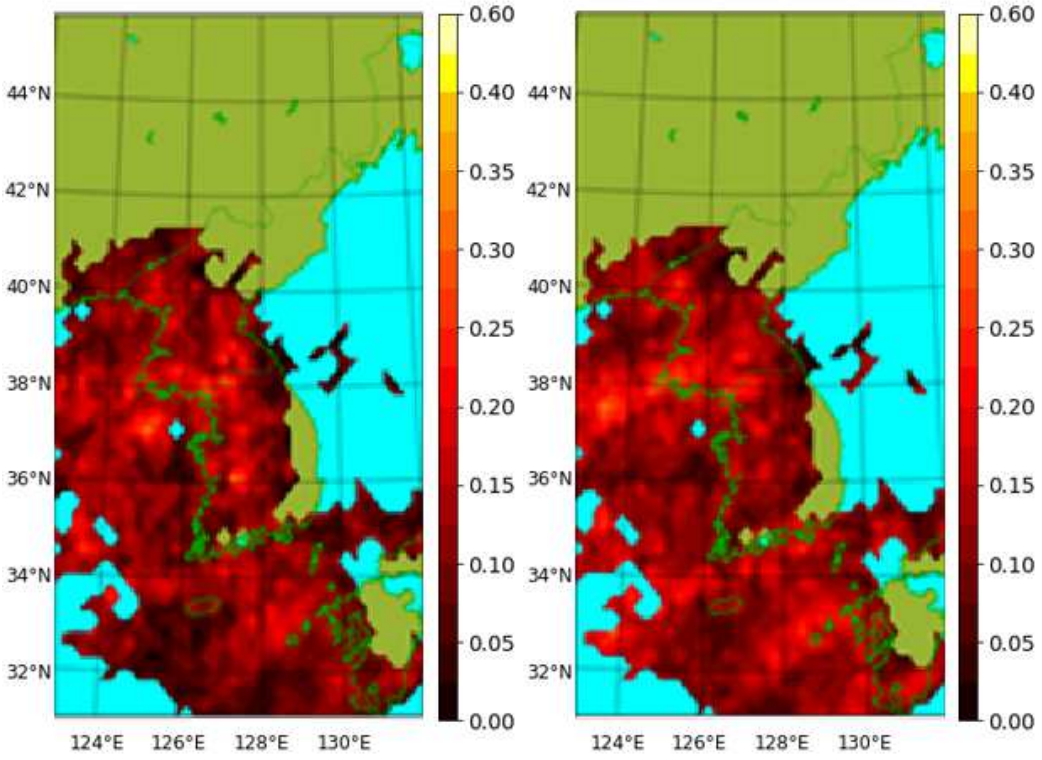
Extended Temporal and Spatial Scales

There were studies (Clark et al., 2010; Lynn et al., 2015) showing that increasing the target radius of lightning forecast can lead to higher ETS values. Therefore, we extended the current temporal and spatial scales to improve the predictive performance of the proposed method. Table 3.6 gives the result of extending the forecast intervals from 3 h to 6 h. With this extended forecast intervals, the 9 h lightning forecast predicts whether or not lightning occurs during the 9–15 h interval after the forecast time. All meteorological variables for that interval are used for training and testing. There was no significant differences in terms of POD, but 6-hour forecast intervals produced about 7 % less false alarms, achieving about 0.04 higher ETS values than 3-hour forecast intervals. Figure 3.8 compares the ETS values for 3-hour and 6-hour forecast intervals. The ETS values for 6-hour intervals are slightly better than those for 3-hour intervals.

We extended the spatial scale by growing the 0.25° grid intervals at latitude and longitude to 0.50° and 0.75°. All meteorological variables belonging to each extended grid are used for

Table 3.6: Performance of the proposed method at 6-hour intervals

Lead time	TP	FP	FN	TN	POD	FAR	ETS
09 h	7,836	34,722	13,645	793,649	0.3648	0.8159	0.1226
15 h	9,659	39,624	12,788	787,781	0.4303	0.8040	0.1375
21 h	7,526	34,541	14,037	793,748	0.3490	0.8211	0.1174
27 h	8,314	38,299	14,191	789,048	0.3694	0.8216	0.1188
Average					0.3784	0.8157	0.1241
Standard deviation					0.0357	0.0082	0.0092



(a) ETS for 3-hour forecast intervals.

(b) ETS for 6-hour forecast intervals.

Figure 3.9: Performance comparison of 3-hour and 6-hour forecast intervals

Table 3.7: Comparison of predictive performance by the grid intervals for latitude and longitude

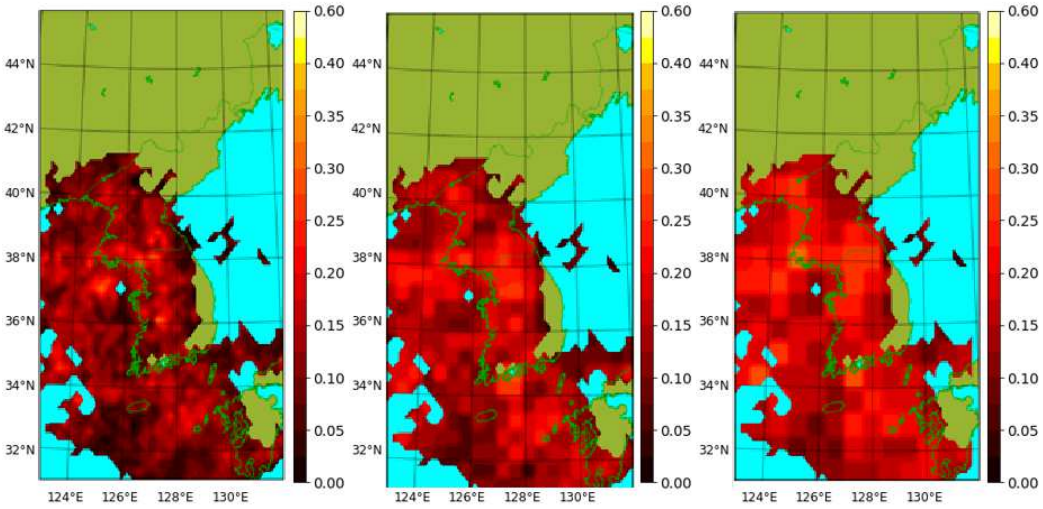
Lead time	POD			FAR			ETS		
	0.25°	0.50°	0.75°	0.25°	0.50°	0.75°	0.25°	0.50°	0.75°
09 h	0.4507	0.4508	**0.4597**	0.8696	0.8026	**0.7433**	0.1007	0.1388	**0.1676**
12 h	0.2922	0.3218	**0.3392**	0.8979	0.8340	**0.7744**	0.0717	0.1047	**0.1301**
15 h	0.4206	0.4383	**0.4460**	0.8817	0.8219	**0.7668**	0.0910	0.1255	**0.1520**
18 h	0.4451	0.4622	**0.4750**	0.8575	0.7911	**0.7340**	0.1081	0.1455	**0.1731**
21 h	0.3988	0.4350	**0.4540**	0.8773	0.8132	**0.7527**	0.0918	0.1297	**0.1605**
24 h	0.2594	0.2875	**0.3106**	0.9052	0.8419	**0.7824**	0.0646	0.0957	**0.1205**
27 h	0.3748	0.3928	**0.4023**	0.8856	0.8288	**0.7764**	0.0855	0.1161	**0.1394**
30 h	0.3871	0.4189	**0.4351**	0.8708	0.8080	**0.7510**	0.0944	0.1290	**0.1557**
Average	0.3786	0.4009	**0.4152**	0.8807	0.8177	**0.7601**	0.0885	0.1231	**0.1499**
SD*	0.0692	0.0636	0.0601	0.0155	0.0171	0.0174	0.0144	0.0168	0.0184

* Standard deviation.
The best values for each performance criterion are shown in bold type.

training and testing. Table 3.7 presents the predictive performance by the grid intervals of latitude and longitude. As grid intervals increased at latitude and longitude, the values of all the performance criteria improved. Figure 3.10 shows the ETS values by the grid intervals, with the highest values at 0.75°. Extending the temporal and spatial scales reduced the resolution of lightning forecast, but improved forecast skill scores.

Undersampling Ratio

Experimental results with various undersampling ratios are presented. Without undersampling, the training process of SVMs did not terminate within a week. Therefore, we reduced the number of grids in the training data. Only 430 grids with more than 10 lightning in July and August each year were used as the training data. Figure 3.11 shows the heat map representing the frequency of lightning during the target period, and Table 3.8 gives the result of SVMs for different undersampling ratios. Since the training data is severely imbalanced, lightning could not be predicted at all without undersampling. Removing many



(a) 0.25° grid intervals. (b) 0.50° grid intervals. (c) 0.75° grid intervals.

Figure 3.10: ETS values by the grid intervals of latitude and longitude

non-lightning instances improves the POD, but tends to increase FAR. Therefore, it is necessary to balance the POD and FAR, and the highest ETS was achieved at the 1:5 ratio. In addition to improving predictive performance, undersampling also significantly reduced training time: at the 1:5 ratio, training time was reduced from 8,127 s to 86 s.

3.3.4 Discussions

It is not easy to forecast lightning activities due to its chaotic characteristics. We applied machine learning techniques to the 113 variables of the ECMWF short-range weather forecasts around the Korean Peninsula, and used undersampling to alleviate the imbalanced data and to speed up training processes. In the lightning prediction, the performance of SVMs was better than that of random forests, and undersampling and the extended temporal and spatial scales improved the predictive performance of the SVMs.

It is also important to predict how many times lightning will occur since it is a good

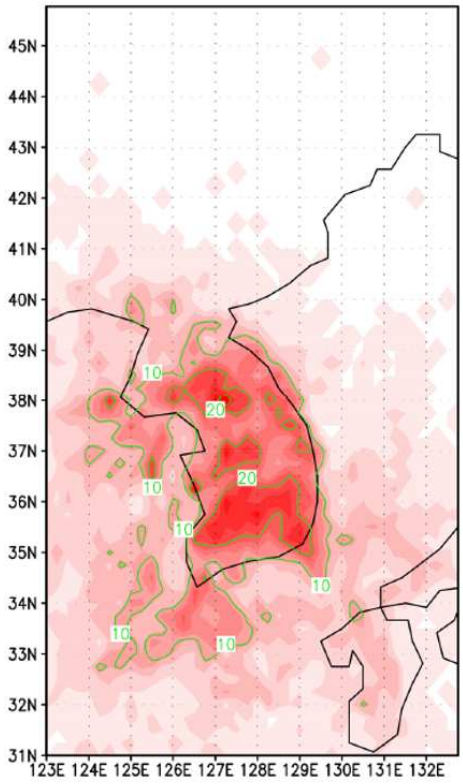


Figure 3.11: Frequency of lightning activities during the target period

Table 3.8: Results of SVMs on 9 h lightning forecasts for different undersampling ratios

Ratio	TP	FP	FN	TN	POD	FAR	ETS	Training (s)	Testing (s)
N/A*	0	0	2,677	103,963	0.0000	NaN	0.0000	8,127.39	0.40
1:1	2,158	23,405	519	80,558	0.8061	0.9156	0.0596	9.70	0.55
1:2	2,271	23,427	406	80,536	**0.8483**	0.9116	0.0639	20.17	0.52
1:3	1,650	10,069	1,027	93,894	0.6164	0.8592	0.1089	43.10	0.58
1:4	1,443	7,631	1,234	96,332	0.5390	0.8410	0.1206	65.09	0.58
1:5	1,303	6,259	1,374	97,704	0.4867	0.8277	**0.1273**	85.87	0.55
1:6	1,095	5,035	1,582	98,928	0.4090	0.8214	0.1245	127.90	0.57
1:7	675	2,638	2,002	101,325	0.2521	**0.7963**	0.1131	147.30	0.56

* Undersampling was not performed.
The best values for each performance criterion are shown in bold type.

indicator of severe weather conditions. In our preliminary work, regression functions such as support vector regression showed promising results, however, they were quite slow compared to the classifiers that we used in this section. Our future work aims to further improve both the computation time and the prediction quality of the lightning forecast models that use regression functions.

Chapter 4

Discretization Techniques

In automatic weather stations, wind direction is usually expressed in degrees. Representing wind direction by continuous values can cause serious problems, depending on the learning algorithm. For example, the difference between 0° and 359° is much smaller than that between 0° and 180°, which can degrade the performance of instance-based learning algorithms such as k-nearest neighbors. Discretization is the process of converting continuous features into nominal ones before model construction. The wind direction, for example, can be expressed as north, east, west, or south wind instead of continuous values through discretization.

There are various discretization methods ranging from a simple method of discretizing with regular intervals to a supervised one of determining the number of intervals and setting discretization boundaries automatically based on class labels of training instances. In this chapter, we propose a selective discretization scheme (Moon et al., 2019), and describe the minimum description length discretization.

Table 4.1: Contingency table for the discretized intervals of hourly precipitation

	$(-\infty, 2\,\text{mm}]$	$(2\,\text{mm}, 5\,\text{mm}]$	$(5\,\text{mm}, 10\,\text{mm}]$	$(10\,\text{mm}, 14\,\text{mm}]$	$(14\,\text{mm}, \infty)$
Advisory[a]	65	61	60	94	77
Non-advisory[b]	41,949	4,714	1,363	632	139
Advisory ratio[c]	0.002	0.013	0.042	0.129	0.356

[a] The advisory criterion for heavy rainfall will be met within the next three hours.
[b] The advisory criterion for heavy rainfall will not be met within the next three hours.
[c] The proportion of advisory instances in each column.

4.1 Selective Discretization

In machine learning, discretization is the process of converting continuous attributes to nominal ones. Many studies have reported that learning algorithms can benefit from the discretization due to the enhanced learning speed and predictive accuracy (Dougherty et al., 1995; Liu et al., 2002); however, information loss is inevitable in the discretization process (Jin et al., 2009), which may degrade the performance of specific learning algorithms. Therefore, we present the selective discretization (Moon et al., 2019) that selectively discretizes attributes to prevent information loss caused by inappropriate discretization of specific numeric attributes.

As an example, a contingency table for the discretized intervals of hourly precipitation is shown in Table 4.1. The higher value of precipitation indicates the higher chance of satisfying the advisory criterion for heavy rainfall. After the discretization, however, the difference between two values within the same interval (e.g., 5 mm and 10 mm) will be ignored; thus, we present a selective discretization method that performs discretization only on the attributes of which the numerical values are not critical within each interval.

Let $T = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ be a training set containing N instances in which $\mathbf{x}_n$ is the M-tuple of attribute values (each instance has M attributes) and y_n is the class label of the n-th instance. When there are K classes, y_n has values from 0 to $K-1$. Discrete

Table 4.2: Contingency table for the discretized intervals of temperature

	$(-\infty, 8.5\,°\mathrm{C}]$	$(8.5\,°\mathrm{C}, 19.2\,°\mathrm{C}]$	$(19.2\,°\mathrm{C}, 23.0\,°\mathrm{C}]$	$(23.0\,°\mathrm{C}, \infty)$
Advisory[a]	0	343	286	8
No advisory[b]	20,152	18,638	7,651	2,464
Advisory ratio[c]	0	0.018	0.036	0.003

[a] The advisory criterion for heavy rainfall will be met within the next three hours.
[b] The advisory criterion for heavy rainfall will not be met within the next three hours.
[c] The proportion of advisory instances in each column.

intervals induced from a discretization method D by discretizing the m-th attribute of T is denoted by $D(T, m) = \{(d_v, d_{v+1}]\}_{v=1}^{V}$, where V is the number of resulting intervals. Then $C(T, m, D) = (c_{kv})$ is a contingency table derived from $D(T, m)$, where c_{kv} denotes the number of the instances whose class labels are k and values of the m-th attribute fall into the v-th interval.

Definition 1. Given a contingency table $C(T, m, D)$, a class label k is *monotonically predictable* on the m-th attribute if the sequence of $(c_{kv}/\sum_h c_{hv})$ is either monotonically increasing or monotonically decreasing for all $1 \leq v \leq V$.

For example, the class *advisory* in Table 4.1 is monotonically predictable on Precipitation (1) since the advisory ratio $(c_{0v}/\sum_h c_{hv})$ is monotonically increasing for all v: $0.002 < 0.013 < 0.042 < 0.129 < 0.356$. More precipitation leads to a higher probability of meeting the advisory criterion for heavy rainfall. The loss of numerical information is not desirable in predicting class labels that are monotonically predictable. On the other hand, Table 4.2 shows the contingency table for the discretized intervals of temperature. The advisory ratios does not increase or decrease monotonically as temperature increases. In this case, the discretized value was used instead of the numerical value of the attribute.

The numerical value of monotonic attributes is important to all class labels. It is easy to see that if one class is monotonically predictable, the other one is also monotonically

predictable in binary classification problems.

Selective discretization first converts all continuous attributes to discrete ones, determines which attributes are monotonic, and undo the discretization of monotonic attributes. Selective discretization scheme S is a binary string of length M, where the m-th character denotes whether the m-th attribute is monotonic (1) or not (0). Monotonic attributes are used as numeric attributes without discretization.

The pseudocode for selective discretization is shown in Figure 4.1. Since the number of intervals cannot exceed the number of instances, the time complexity of the selective discretization excluding the execution time of the discretization method D is $\mathcal{O}(MN)$, where M is the number of attributes, N is the number of instances, and the number of class labels is assumed to be constant.

The rationales for the selective discretization are that (a) it is better to let classifiers handle monotonic attributes directly than to give them discretized ones since the discrete values of monotonic attributes cannot fully utilize information between numerical values and class labels; and (b) discretizing attributes that have a nonlinear relationship with class labels can help linear models that have difficulty in processing nonlinear attributes.

4.2 Minimum Description Length Discretization

The minimum description length (MDL) discretization (Fayyad and Irani, 1993) is an entropy-based supervised discretization method. The MDL method defines the class entropy of an instance set T with K class labels as:

$$\mathrm{H}(T) = -\sum_{k=0}^{K-1} p(T,k) \log p(T,k),$$

where $p(T,k)$ is the proportion of instances in T of which class labels are k. The class entropy measures the amount of information needed to specify the classes in T.

Algorithm 1: Selective discretization

Input: a discretization method D, a training set T, and a set of attributes $A = \{A_1, A_2, \ldots, A_l\}$ associated with T

Output: selective discretization scheme S

```
1  set S to ∅;
2  set m to the number of class labels in T;
3  for k ← 1 to l do
4      if A_k is a continuous attribute then
5          set monotonic to true;
6          set J_k to D(T, A_k);
7          set C_k to the contingency table constructed by T and J_k;
8          set n to the number of intervals in J_k;
9          for j ← 1 to n do
10             set s_j to the sum of the j-th column in C_k;
11         for i ← 1 to m do
12             for j ← 1 to n do
13                 set c_ij to the (i, j) entry of C_k;
14                 set r_ij to c_jk/s_j;
15             for j ← 2 to n − 1 do
16                 if r_ij is not between r_ij−1 and r_ij+1 then
17                     set monotonic to false;
18         if monotonic is true then
19             add (A_k, ∅) to S;
20         else
21             add (A_k, J_k) to S;
22     else
23         add (A_k, ∅) to S;
```

Figure 4.1: Pseudocode for selective discretization

Given an interval boundary d on the m-th attribute of the instance set T, let T_1 be the subset of instances in T with the values of the m-th attribute is less than or equal to d and $T_2 = T - T_1$. The class information entropy of the partition induced by d, $\mathrm{H}(m, d; T)$, is defined as:

$$\mathrm{H}(m, d; T) = \frac{|T_1|}{|T|}\mathrm{H}(T_1) + \frac{|T_2|}{|T|}\mathrm{H}(T_2).$$

A binary discretization for the m-th attribute is determined by selecting the boundary d_{min} for which $\mathrm{H}(m, d_{min}; T)$ is minimal among all the possible interval boundaries. This binary discretization is applied recursively to both of the partitions induced by d_{min} until the stopping criterion based on the MDL principle is met. A detailed account of the stopping criterion can be found in Fayyad and Irani (1993).

Liu et al. (2002) compared eight discretization methods using eleven benchmark data sets. Among the methods, the MDL method had the highest classification accuracy. The default discretization method used by the selective discretization scheme is the MDL method.

4.3 Case Study: Heavy Rainfall Forecast

The purpose of an early warning system (EWS) is to issue warning signals prior to extreme events. Extreme weather events, however, are hard to predict due to their chaotic behavior. This section suggests a method for an effective EWS for very short-range heavy rainfall with machine learning techniques. The EWS produces a warning signal when it is expected to reach the criterion for a heavy rain advisory within the next 3 hours. Meteorological data obtained from automatic weather stations are preprocessed by the selective discretization and principal component analysis. As a classifier, logistic regression is used to predict whether or not a warning is required.

4.3.1 Introduction

An EWS produces a warning signal before a dangerous event occurs so that we can prepare for the event. Alfieri et al. (2012) reviewed operational EWSs for water-related hazards such as floods and landslides in Europe. An EWS for heavy rain using meteorological radar and pluviometers was successfully operated in Rio de Janeiro (Heffer, 2013). In Japan, an EWS for heavy rain using multi-parameter phased array weather radar is tested for use in the Tokyo 2020 Olympics (Kobayashi, 2018). Kim and Yoon (2016) analyzed spatiotemporal patterns of heavy rain to predict whether or not it will occur within three hours at each automatic weather stations (AWSs) in South Korea. In this section, an EWS for heavy precipitation using meteorological data from AWSs is proposed and its performance is measured by various criteria.

A short-range weather forecast within the next 3 hours is often referred to as *nowcasting* (Glossary of Meteorology, 2019a), and it plays an important role in the crisis management of natural disasters. Numerical weather prediction is a traditional method to predict precipitation. Given the current weather conditions, it uses mathematical models to simulate the atmosphere and forecasts the future state of the weather. This method, however, is not appropriate for regional nowcasting because of a spin-up problem and a low spatial and temporal resolution (Mecklenburg et al., 2000). An alternative approach is needed to complement the numerical predictions on a smaller spatial and temporal scale.

Recently, machine learning techniques have been used to forecast rainfall with the progress in the field of pattern recognition and artificial intelligence. Classifiers can be used in the rain/no-rain classification (Liu et al., 2001; Meyer et al., 2016) or the prediction of heavy rainfall (Lee et al., 2012; Seo et al., 2014), and regression functions can be used to predict the amount of precipitation (Toth et al., 2000; Ramírez et al., 2005; Hong, 2008; Chattopadhyay and Chattopadhyay, 2010; Nastos et al., 2014) and to detect anomalies in meteorological data (Lee et al., 2018). In particular, regression functions based on artificial

neural networks are prevalently used to predict hydrological time series data. For example, a nonlinear autoregressive network with exogenous inputs (NARX) was used to forecast flood (Chang et al., 2014; Nanda et al., 2016; Chang et al., 2018) and groundwater levels (Wunsch et al., 2018), an adaptive network-based fuzzy inference system (ANFIS) was used in real-time reservoir operation model (Hsu et al., 2015), flood forecasting (Chang and Tsai, 2016) and streamflow forecasting (Yaseen et al., 2017), self-organizing map (SOM) was used to forecast monthly precipitation (Rivera et al., 2012), and long short-term memory (LSTM) and gated recurrent unit (GRU) are used to predict combined sewer overflow (Zhang et al., 2018).

The performance of a machine learning algorithm is often measured by the accuracy, or equivalently, the error rate. Accordingly, it is common to evaluate classifiers by the rate of correct classification and regression functions by the mean squared error. When data set is highly imbalanced, however, the overall prediction accuracy may be misleading (Chawla et al., 2002; He and Garcia, 2009; Liu et al., 2009; Su et al., 2006; Sun et al., 2007). For example, Seoul, the capital of South Korea, needed heavy rain advisories for 58 hours while it did not for the other 52,608 hours from 2007 to 2012. A learning algorithm could decide not to issue heavy rain advisories at all so that it can achieve 99.89 % accuracy, which is meaningless. The EWS needs a proper performance criterion other than the accuracy.

We investigate the possibility of employing machine learning techniques in constructing an early warning system for heavy rainfall with a lead time of 3 hours. Meteorological data are preprocessed by the selective discretization and principal component analysis (PCA), and logistic regression is used as a classifier. A comparative analysis was conducted on various classifiers with a conventional discretization method, the selective discretization, PCA, and their combinations.

As far as we know, our selective discretization which applies discretization to only a few selected input variables was the first attempt and it could help to predict very short-range

Table 4.3: Input variables for the EWS

No.	Variable name	Description
1	Date	Day count from January 1 ([1, 365 or 366])
2	Time	Hour value of 24-hour clock ([1, 24])
3	Wind direction	Average wind direction for the last 10 minutes (°)
4	Scalar wind speed	Average wind speed for the last 10 minutes ($\mathrm{m\,s^{-1}}$)
5	Vertical wind speed	Average magnitude of the North-South component for the last 10 minutes ($\mathrm{m\,s^{-1}}$)
6	Horizontal wind speed	Average magnitude of the East-West component for the last 10 minutes ($\mathrm{m\,s^{-1}}$)
7	Temperature	Average temperature for the last 1 minute (°C)
8	Humidity	Average humidity for the last 1 minute (%)
9	Atmospheric pressure	Average atmospheric pressure for the last 1 minute (hPa)
10	MSLP	Average mean sea level pressure for the last 1 minute (hPa)
11	Rain sensor	Indication of whether or not it is raining (0 or 1)
12-23	Precipitation (h)	Amount of precipitation for the last h (1 to 12) hours (mm)

heavy rainfall. It is expected that the comparative experiments of the various techniques in this section will be helpful in constructing a system for predicting various meteorological elements using machine learning techniques.

4.3.2 Early Warning System

Input Variables

An EWS takes input from its connected AWS. Regional meteorological data such as wind, temperature, humidity, atmospheric pressure and the amount of precipitation are provided to the EWS every hour. The input variables used in this section are listed in Table 4.3.

Classifier

We use logistic regression to predict heavy rainfall. Logistic regression is a classifier that is a type of regression analysis for binary classification. It assumes a linear relationship between the log odds of the dependent variable and the independent variables. Logistic regression

is used widely in many areas including the medical and social science, engineering, and econometrics. Logistic regression uses a logistic function to predict binary outcomes from continuous ones produced by regression analysis. Logistic function is a sigmoid function with the equation:

$$f(x) = \frac{1}{1 + e^{-x}},$$

where e is the base of the natural logarithm. The graph of the logistic function is shown in Fig. 4.2. Logistic regression substitutes x with a linear function in the feature space such that:

$$x = \alpha_0 + \alpha_1 a_1 + \alpha_2 a_2 + \cdots + \alpha_M a_M,$$

where M is the number of attributes, α_0 is the intercept coefficient, α_m is the m-th regression coefficient, and a_m is the value of the m-th attribute for all $1 \leq m \leq M$. The value of the logistic function can be interpreted as the probability that the criterion for a heavy rain advisory will be met within the next 3 hours. The regression coefficients are generally estimated by the maximum likelihood method (Hosmer et al., 2013). In this section, the ridge estimator (Cessie and Houwelingen, 1992) is used to prevent overfitting and unstable estimates, and nominal attributes are converted to binary numeric attributes just as in PCA.

Functionality

A heavy rain advisory is issued when the precipitation for 6 hours is expected to be more than 70 mm or the precipitation for 12 hours to be over 110 mm. In this section, the purpose of the EWS is to issue a warning signal when the heavy rain advisory condition is likely to be satisfied within the next 3 hours. Let $\Pi([t_1, t_2])$ be the amount of precipitation in millimeters between time t_1 and t_2, and for integer h, $\oplus(t, h)$ be the time h hours later from

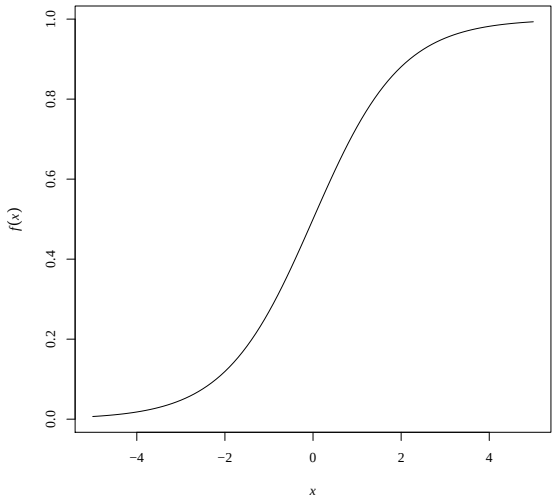


Figure 4.2: Shape of the logistic function, $f(x) = \frac{1}{1+e^{-x}}$

the time t if h is positive, or earlier if negative. At prediction time t_0, the EWS should issue a warning signal if and only if $\Pi([\oplus(t_0, -3), \oplus(t_0, 3)]) \geq 70$ or $\Pi([\oplus(t_0, -9), \oplus(t_0, 3)]) \geq 110$. Figure 4.3 provides an alternative representation of the warning criterion for the EWS.

Architecture

In South Korea, there are over 600 AWSs that measure and report weather conditions automatically. As shown in Figure 4.4, the stations are located all over the country and provide real-time meteorological data. A dedicated EWS for heavy rainfall nowcasting is constructed for each station. The architecture of an EWS is depicted in Figure 4.5. The EWS is first trained on the meteorological database of a specific region: data preprocessing methods and the classifier of the EWS are trained for heavy rainfall nowcasting. After training, the EWS takes real-time input from its connected AWS, and produces an output, whether or not a warning is required. The AWS updates the meteorological database with

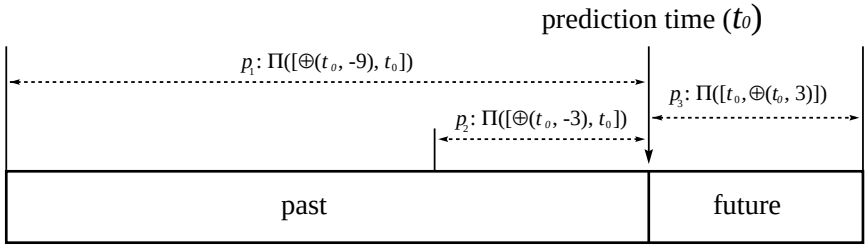


Figure 4.3: Warning criterion of the EWS for very short-range heavy rainfall

At prediction time, p_1 is the amount of precipitation in the last 9 hours, and p_2 is the amount of precipitation in the last 3 hours. The EWS should issue a warning signal if and only if p_3, the amount of precipitation within the next 3 hours, is over $110 - p_1$ or $70 - p_2$.

recent weather data periodically, and the EWS can be retrained with the renewed database.

Performance Criteria

A confusion matrix is typically used to visualize the performance of machine learning algorithms. The confusion matrix of the EWS is illustrated in Table 4.4. In the matrix, true positive (TP) is the number of correct warnings, and false positive (FP) is the number of incorrect warnings. In contrast, true negative (TN) is the number of the correct predictions that did not issue a warning, and false negative (FN) is the number of the incorrect predictions that failed to issue a warning when it was needed.

In pattern recognition, many performance criteria are the functions of the confusion matrix. Precision and recall are commonly used metrics to quantify the performance of learning algorithms (Forman, 2003; Caruana and Niculescu-Mizil, 2006; Davis and Goadrich, 2006), and they are defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

Precision denotes the percentage of warning signals that are correct, but does not take into account FN. Recall is the percentage of advisory instances that are correctly classified, but

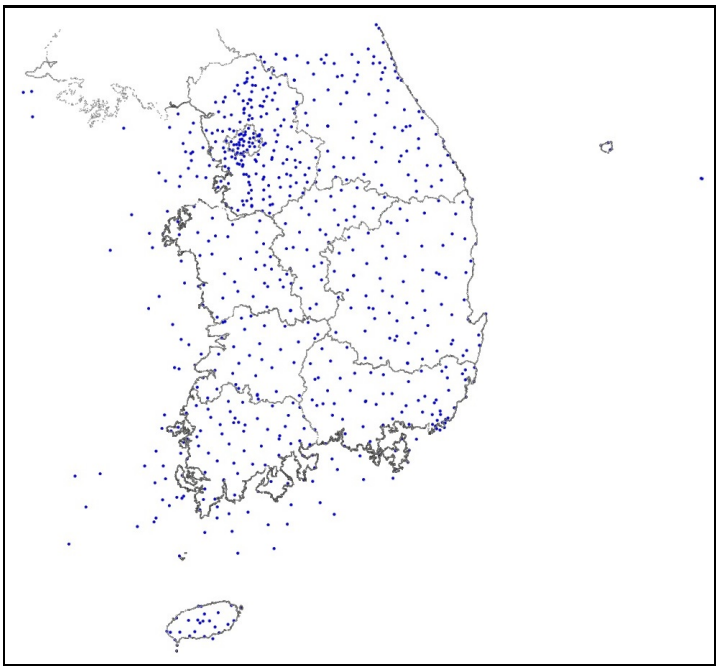


Figure 4.4: Locations of automatic weather stations in South Korea

false alarms are not taken into account. Since there is an inverse relationship between the two measures, these metrics need to be considered together. For example, issuing warning signals all the time achieves recall of 100 %, whereas it reduces precision significantly in most cases. F-measure, which is the harmonic mean of precision and recall, provides a balanced measure:

$$\text{F-measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Since F-measure is often used as the ultimate measure of performance of classifiers (Forman, 2003), it is also used to measure the performance of learning algorithms in imbalanced classification (Sun et al., 2007; Liu et al., 2009).

Rainfall forecasts are often verified by the probability of detection (POD) and the false

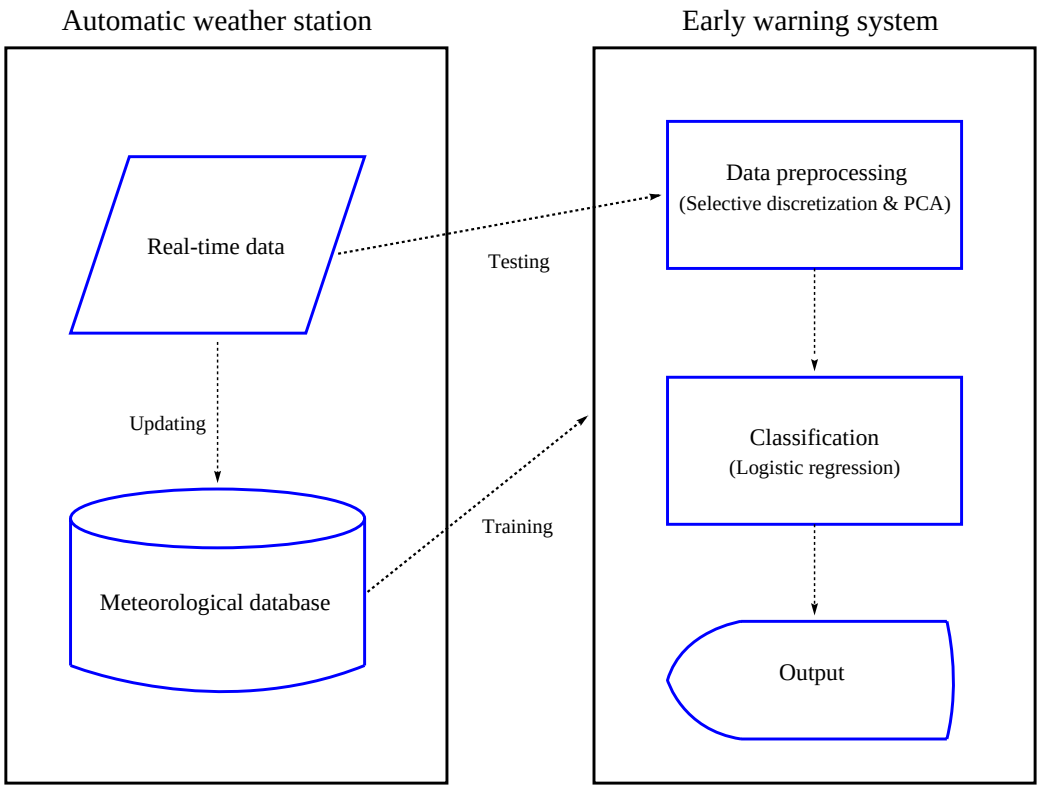


Figure 4.5: Architecture of the EWS for very short-range heavy rainfall

alarm rate (FAR), which are closely related to precision and recall. The POD is equivalent to recall, and the FAR equals to $1 - \text{Precision}$. The threat score (TS) provides a more balanced measure than the POD and the FAR by taking account into both FP and FN:

$$\text{TS} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}.$$

Qian et al. (2016) used the TS for assessing the detection of heavy precipitation area in China, and used the POD and the FAR for providing a better understanding of a given TS value. However, the TS is sensitive to climatological frequency of events and is not appropriate for the forecasts of rare events due to the number of the correct predictions that occurred by random chance. The equitable threat score (ETS) adjusts the TS by excluding the expected number of correct forecasts that happened by chance:

$$\text{ETS} = \frac{\text{TP} - \alpha}{\text{TP} + \text{FP} + \text{FN} - \alpha},$$

where

$$\alpha = \frac{(\text{TP} + \text{FP}) \cdot (\text{TP} + \text{FN})}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}.$$

The ETS is often used to assess the quality of forecasts of rare events such as precipitation above a large threshold (Jolliffe and Stephenson, 2003), and Meyer et al. (2016) used the POD, FAR, TS and ETS for the validation of rainfall area predictions.

In this section, we use F-measure and ETS to measure the performance of EWSs since the F-measure is suitable for evaluating the performance of classifiers and the ETS is suitable for assessing the quality of forecasts. The two criteria are appropriate for forecasting rare events, as opposed to accuracy which is highly affected by TN, which can be very large in imbalanced classification. Higher scores of both measures mean better performance, and they

Table 4.4: Confusion matrix for the EWS

	Warning was issued	Warning was not issued
Advisory[a]	True positive (TP)	False negative (FN)
No advisory[b]	False positive (FP)	True negative (TN)

[a] The advisory criterion will be met within the next 3 hours.
[b] The advisory criterion will not be met within the next 3 hours.

equal to 1 when all predictions are correct. Since the two criteria treat FP and FN equally, however, they can be inappropriate when one is more critical than the other. Refer to Jolliffe and Stephenson (2003) and Wilks (2011) for general guidance on forecast verification.

4.3.3 Experiments

We conducted experiments to evaluate the performance of the proposed method, in which selective discretization and PCA are applied to preprocess training data, and logistic regression is then used to predict heavy rainfall.

Experimental Setup

We used hourly meteorological data from 652 AWSs in South Korea from 2007 to 2012. The total amount of the data is approximately 3 GB. The average number of instances for each station is 46,200, while the average number of *advisory* instances, which will satisfy the criterion for a heavy rain advisory within the next 3 hours, is only 36. A classifier that never predicts a warning achieves 99.9 % accuracy. Instances that met the heavy rainfall criterion without future precipitation, i.e., the advisory instances with the value of *Precipitation (3)* greater than 70 mm or the value of *Precipitation (9)* greater than 110 mm, were excluded from the experiment. Instances with missing values in more than three attributes were also excluded.

In the evaluation of time series forecasting, it is very common to use the last block

evaluation, which uses the first part of the time series as a training set and the rest as a testing set; however, cross-validation is a more robust method than the last block evaluation for the model selection in time series forecasting (Bergmeir and Benitez, 2012). We performed stratified 3-fold cross validations so that each fold contains roughly the same number of *advisory* instances. The cross validation process was repeated for 30 times with different random samples of 3 folds.

The no free lunch (NFL) theorem (Wolpert and Macready, 1997) states that there is no single learning algorithm that works best on all purposes; thus, we tested various types of classifiers to determine which one is better suited for very short-range heavy rain prediction. The classifiers used in the experiments were logistic regression, artificial neural network (ANN), 1-nearest neighbor (1-NN) (Aha and Kibler, 1991), C4.5 (Quinlan, 1993), random forests (Breiman, 2001), LIBSVM (a library for support vector machines) (Chang and Lin, 2011), SMO (sequential minimal optimization) (Platt, 1999), and RIPPER (repeated incremental pruning to produce error reduction) (Cohen, 1995).

Multilayer perceptron (MLP) is used as a feedforward ANN. The MLP uses backpropagation to train the network (Rumelhart et al., 1986). The universal approximation theorem (Hornik, 1991) states that feedforward networks with as few as a single hidden layer are universal approximators under some general conditions. The 1-NN is an instance-based learning algorithm that assigns an instance to the class of its closest neighbor in the attribute space. In the large sample case, the error rate of 1-NN is less than twice the Bayes error rate which is the minimum probability of error given the distribution of the data (Cover and Hart, 1967). The C4.5 is a decision tree learner that builds a decision tree using the concept of information entropy. It generates a tree by recursively choosing the attribute that best differentiates instances of the training set at each node of the tree. To avoid overfitting, pruning is carried out from leaves to the root. The C4.5 was selected for the top 10 algorithms in data mining (Wu et al., 2007). Random forests is an ensemble

learning technique that constructs a number of decision trees on various subsamples of the training set and takes a majority vote to classify an instance. It can be seen as a bagging algorithm (Breiman, 1996) for decision trees that uses only a random subset of attributes for splitting each node. The bagging improves the predictive accuracy and helps to avoid overfitting. LIBSVM and SMO are support vector machines (SVMs) (Burges, 1998), which try to seek the hyperplane that separates training instances according to their class labels with the largest margin. The maximum-margin hyperplane is expected to have good generalization on unseen data. The SMO uses the algorithm of Platt (1999) and the LIBSVM uses the SMO-type method proposed in Fan et al. (2005) to train SVMs. The RIPPER is a classifier that learns propositional rules and is designed to perform efficiently on large noisy datasets. It uses repeated grow-and-simplify approach to build a rule set. Cohen (1995) showed that RIPPER was generally better than C4.5 rules (Quinlan, 1993) which derives rules from a decision tree.

The Waikato Environment for Knowledge Analysis (WEKA) package (version 3.8.1) due to Hall et al. (2009) was used to implement the classifiers, the MDL method, and PCA. The selective discretization method was implemented in C#. Default settings were used for all programs except ANN and LIBSVM. The training time of the ANN was reduced from 500 to 50 for speedup, and the normalization and the probability estimation options of the LIBSVM were turned on to improve predictive accuracy.

Comparative Analysis

The performance of the proposed method was compared to various EWS models. The results are summarized in Table 4.5. For each measure, the average of 30 trials were computed, but the stratified 3-fold cross validation was performed once for ANN and 1-NN due to their long computation time. As stated previously, precision and recall have an inverse relationship, and F-measure and ETS are more balanced measures. All measures were computed over

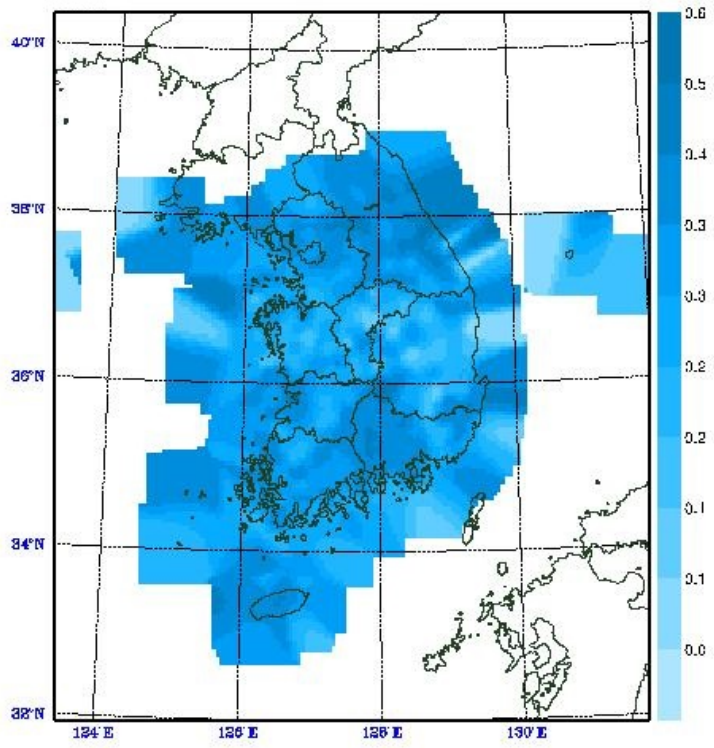


Figure 4.6: Heat map displaying the ETS values by the proposed method

A darker color indicates a higher value of ETS for the corresponding region.

all stations for each run, and the results were averaged over 30 runs. When both selective discretization and PCA were used together, the performances of all classifiers were improved in terms of F-measure and ETS. The proposed method, which is the logistic regression with the selective discretization and the PCA, achieved the highest F-measure and ETS. The ETS values of the proposed method is shown by the heat map in Figure 4.6. The performance of the proposed method was not affected so much by the locations of the AWSs.

The performances of the classifiers for each station are compared in Table 4.6. For each classifier, the EWS model with the highest F-measure and ETS was selected. Each item in

Table 4.5: Comparison of the performance of various EWS models via stratified 3-fold cross validations

Classifier	Discretization	PCA	Precision	Recall	F-measure	ETS
Logistic	No	No	0.4888	0.3627	0.4164	0.2445
	MDL	No	0.4485	0.3603	0.3996	0.2493
	SD	No	0.5088	**0.3942**	0.4442	0.2852
	No	Yes	0.5445	0.3541	0.4291	0.2729
	MDL	Yes	0.4989	0.3730	0.4268	0.2710
	SD	Yes	**0.5590**	0.3909	**0.4601**	**0.2985**
C4.5	No	No	0.4680	0.3250	0.3836	0.2370
	MDL	No	**0.5263**	0.2511	0.3402	0.2047
	SD	No	0.4943	0.3112	0.3819	0.2357
	No	Yes	0.4619	0.2953	0.3602	0.2194
	MDL	Yes	0.4663	0.3086	0.3714	0.2277
	SD	Yes	0.4673	**0.3318**	**0.3881**	**0.2404**
Forests	No	No	**0.6191**	0.2881	0.3932	0.2445
	MDL	No	0.5032	**0.3443**	**0.4089**	**0.2567**
	SD	No	0.6173	0.3037	0.4071	0.2553
	No	Yes	0.5426	0.2839	0.3728	0.2288
	MDL	Yes	0.4889	0.3176	0.3850	0.2381
	SD	Yes	0.5762	0.3062	0.3999	0.2496
LIBSVM	No	No	0.5614	0.3128	0.4017	0.2511
	MDL	No	**0.6301**	0.2438	0.3516	0.2131
	SD	No	0.5687	0.3156	0.4060	0.2544
	No	Yes	0.5246	0.3298	0.4050	0.2529
	MDL	Yes	0.4888	0.2707	0.3484	0.2107
	SD	Yes	0.5381	**0.3547**	**0.4275**	**0.2716**
SMO	No	No	**0.6729**	0.2204	0.3321	0.1989
	MDL	No	0.5655	0.3078	0.3986	0.2486
	SD	No	0.6476	0.2843	0.3951	0.2459
	No	Yes	0.6078	0.2779	0.3814	0.2354
	MDL	Yes	0.5140	**0.3421**	0.4108	0.2582
	SD	Yes	0.5950	0.3398	**0.4325**	**0.2757**
RIPPER	No	No	0.3976	0.4006	0.3991	0.2489
	MDL	No	**0.4535**	0.3591	0.4008	0.2503
	SD	No	0.4059	0.3988	0.4023	0.2514
	No	Yes	0.4142	0.3848	0.3990	0.2488
	MDL	Yes	0.4230	0.3892	0.4054	0.2539
	SD	Yes	0.4219	**0.4122**	**0.4170**	**0.2630**
ANN	No	No	0.5558	0.3122	0.3998	0.2496
	MDL	No	0.5015	**0.3762**	0.4299	0.2735
	SD	No	0.5597	0.3621	**0.4397**	**0.2815**
	No	Yes	0.5582	0.2984	0.3889	0.2411
	MDL	Yes	0.5470	0.3478	0.4252	0.2697
	SD	Yes	**0.5721**	0.3409	0.4273	0.2714
1-NN	No	No	0.4324	0.3967	0.4138	0.2605
	MDL	No	**0.5006**	0.3244	0.3937	0.2448
	SD	No	0.4786	**0.4241**	**0.4497**	**0.2897**
	No	Yes	0.3772	0.3627	0.3698	0.2265
	MDL	Yes	0.4256	0.3333	0.3738	0.2295
	SD	Yes	0.4361	0.4132	0.4243	0.2689

'MDL' denotes the discretization by the MDL method
'SD' denotes the selective discretization method.
The best values are indicated in bold type for each classifier.

Table 4.6 contains win-tie-loss information according to the Wilcoxon signed-rank test. The test is a non-parametric alternative to the paired t-test and suitable for comparison of two classifiers (Demšar, 2006). For each comparison, EWS models with the highest F-measure and ETS was selected. These tests were conducted individually on each station. Based on the results of 30 runs of the 3-fold cross validation on a station, it is decided whether or not one classifier performs significantly better than the other at the station. Only statistically significant wins and losses are accepted, and when there is no significant difference in performance, it is counted as a tie. For ANN and 1-NN, the win-tie-loss percentage is calculated on the result of a single run instead of the Wilcoxon signed-rank test. The results showed that the proposed method overall outperforms all the others, and one can see that there is no notable difference between the F-measure and the ETS.

One may wonder why simple logistic regression was superior to other more sophisticated techniques. In fact, the performances of all the techniques were almost the same in terms of accuracy, which is the ratio of the number of correct classifications to the total number of classifications. The accuracies of all the techniques used in this section ranged from 99.90 % to 99.93 %. For example, the accuracy of the proposed method and the SMO model with the lowest ETS in Table 4.5 was almost the same at 99.92 %. The NFL theorem implies that there is a classifier that is appropriate for a particular field, and the logistic regression seems to be suited to the EWS for very short-range heavy rainfall evaluated by F-measure and ETS.

Effects of the Preprocessing Methods

As preprocessing, we selectively discretized input variables. All continuous variables were first discretized by the MDL method and checked to see if they were monotonic. The variables that were not monotonic remained discretized, and the discretization of monotonic variables

Table 4.6: Wilcoxon signed-rank tests on F-measure and ETS with the significance level at 0.01

	Logistic	C4.5	Forests	LIBSVM	SMO	RIPPER	ANN	1-NN
Logistic	-	73-26-1	65-31-4	51-40-9	54-41-5	57-38-5	59-6-35	49-5-46
C4.5	1-26-73	-	9-62-29	6-41-53	8-40-52	1-64-35	25-7-68	27-5-68
Forests	4-31-65	29-61-10	-	13-47-40	13-47-40	14-61-25	35-6-59	31-5-64
LIBSVM	9-39-52	53-40-7	39-46-15	-	26-46-28	31-52-17	44-7-49	41-7-52
SMO	5-41-54	52-39-9	41-46-13	29-45-26	-	34-48-18	46-7-47	41-6-53
RIPPER	5-38-57	35-64-1	25-61-14	18-52-30	18-47-35	-	35-5-60	33-5-62
ANN	35-6-59	69-6-25	59-6-35	49-7-44	47-6-47	60-5-35	-	42-6-52
1-NN	46-5-49	68-5-27	64-5-31	53-6-41	54-6-40	62-5-33	52-6-42	-

Results for F-measure are shown in the upper triangle, and those for ETS are in the lower triangle.
Each item indicates the win-tie-loss percentage of a classifier in that row comparing with a classifier in that column.

were rolled back. Table 4.7 shows the result of selective discretization. The number of intervals of each variable and whether or not the variable is selected in the selective discretization are shown. All values are averaged over 30 trials of stratified 3-fold cross validations. The result indicates that time and wind were not helpful in predicting very short-range heavy rainfall since the number of intervals is close to one. When the number of intervals of a variable is one, the variable always has the same value regardless of the weather conditions. On the other hand, date and temperature were selected to remain discretized over a half of the stations. The selected variables have a nonlinear relationship with the very short-range heavy rainfall. The performance analysis of the selective discretization is shown in Table 4.8. In general, the percentage of stations that were improved by the selective discretization was greater than the percentage of stations that were worsened.

The PCA reduces the dimensionality of input variables by introducing a new coordinate system. In the experiments, the average number of available input variables was 20.8, and it was reduced to 9.6 by the PCA. When the MDL method was used together, the number increased to 22.3, while it decreased to 12.3 in the case that the selective discretization was

Table 4.7: Result of the selected discretization

Variable name	Intervals	Selected (%)
Date	3.93	84.44
Time	1.02	0.02
Wind direction	1.23	3.99
Scalar wind speed	1.36	0.21
Vertical wind speed	1.46	4.86
Horizontal wind speed	1.42	1.78
Temperature	2.53	52.25
Humidity	2.16	4.63
Atmospheric pressure	2.56	1.20
MSLP	2.61	1.19
Precipitation (1)	3.40	0.22
Precipitation (2)	3.63	0.15
Precipitation (3)	3.74	0.14
Precipitation (4)	3.69	0.39
Precipitation (5)	3.69	0.28
Precipitation (6)	3.71	0.39
Precipitation (7)	3.73	0.27
Precipitation (8)	3.73	0.51
Precipitation (9)	3.70	0.62
Precipitation (10)	3.68	0.45
Precipitation (11)	3.66	0.49
Precipitation (12)	3.63	0.49

Table 4.8: Wilcoxon signed-rank tests whether or not the selective discretization significantly improves the predictive performance

Classifier	Discretization	PCA	F-measure	ETS
Logistic	No	No	47-46-7	48-45-7
	MDL	No	56-39-5	55-39-6
	No	Yes	52-39-9	52-38-10
	MDL	Yes	49-44-7	49-44-7
C4.5	No	No	6-82-12	7-81-12
	MDL	No	43-55-2	43-55-2
	No	Yes	32-65-3	32-64-4
	MDL	Yes	26-67-7	26-66-8
Forests	No	No	18-78-4	18-77-5
	MDL	No	10-74-16	10-75-15
	No	Yes	35-62-3	35-62-3
	MDL	Yes	23-66-11	24-65-11
LIBSVM	No	No	22-64-14	23-63-14
	MDL	No	59-36-5	59-36-5
	No	Yes	43-47-10	46-43-11
	MDL	Yes	65-30-5	65-30-5
SMO	No	No	64-35-1	71-27-2
	MDL	No	24-47-29	23-47-30
	No	Yes	59-37-4	62-34-4
	MDL	Yes	41-44-15	40-45-15
RIPPER	No	No	9-87-4	8-86-6
	MDL	No	10-83-7	10-83-7
	No	Yes	24-74-2	24-73-3
	MDL	Yes	23-70-7	23-70-7
ANN	No	No	65-8-27	65-8-27
	MDL	No	51-11-38	51-11-38
	No	Yes	63-11-26	65-9-26
	MDL	Yes	45-13-42	45-13-42
1-NN	No	No	68-5-27	68-5-27
	MDL	No	65-12-23	65-11-24
	No	Yes	72-4-24	72-4-24
	MDL	Yes	64-10-26	64-10-26

Tests were conducted individually on each station with the significance level at 0.01.
For ANN and 1-NN, the win-tie-loss percentage is calculated on the result of a single run.
Each item indicates the win-tie-loss percentage of the EWS model whose discretization method was replaced with the selective discretization comparing with the original model in that row.

used instead. As stated earlier, the number of input variables can be increased by the PCA since the nominal-to-binary filter is applied to nominal variables. The performance analysis of the PCA is shown in Table 4.9. While logistic regression and SVMs benefited much from the PCA, the others did not.

When selective discretization and PCA used together, the overall performances of all classifiers were improved as shown in Table 4.5. The performance analysis of the EWS models that use both methods in the data preprocessing step is presented in Table 4.10. The percentage of improved stations was greater than the percentage of stations that were worsened by both methods, and the Wilcoxon signed-rank test indicates that more than 60 percent of all stations were significantly improved in logistic regression and SMO.

In logistic regression, both the selective discretization and the PCA significantly improved predictive accuracy in terms of F-measure and ETS. The effect of the preprocessing methods on the regions in South Korea is shown in Figure 4.7. One can see that each method helps to predict very short-range heavy rainfall in many regions.

Running Time Analysis

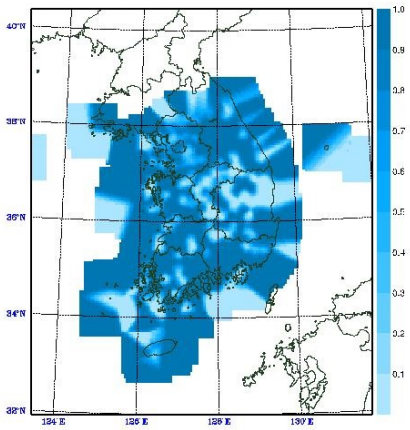
The computation time of the EWS models consists of data preprocessing time, training time, and testing time. The data preprocessing is performed prior to both training and testing; therefore, the data preprocessing time is independent of the classifier used. A running time analysis of the EWS models is shown in Table 4.11.

Although the data processing methods take some additional time, they reduce the total computation time in some cases. For example, the MDL method significantly reduced the training time of decision tree learners by replacing the built-in discretization methods of their own and the PCA reduced the computation time of some classifiers by decreasing the number of input variables. However, the MDL method did not match well with the PCA

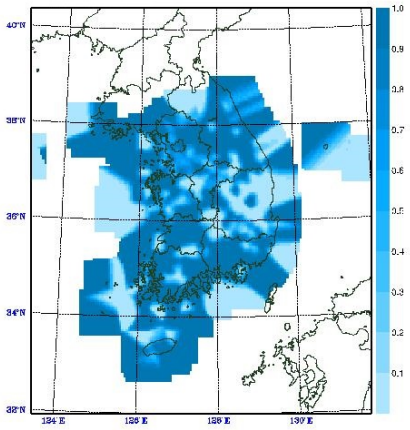
Table 4.9: Wilcoxon singed-rank tests whether or not PCA significantly improves the predictive performance

Classifier	Discretization	PCA	F-measure	ETS
Logistic	No	No	42-41-17	42-41-17
	MDL	No	49-51-0	48-52-0
	SD	No	43-46-11	43-46-11
C4.5	No	No	12-55-33	12-54-34
	MDL	No	37-61-2	38-60-2
	SD	No	24-63-13	24-63-13
Forests	No	No	10-57-33	10-57-33
	MDL	No	0-67-33	0-67-33
	SD	No	15-65-20	15-65-20
LIBSVM	No	No	31-43-26	31-42-27
	MDL	No	30-48-22	30-48-22
	SD	No	45-40-15	46-39-15
SMO	No	No	65-33-2	72-26-2
	MDL	No	25-69-6	26-69-5
	SD	No	63-35-2	63-35-2
RIPPER	No	No	19-62-19	18-63-19
	MDL	No	12-80-8	13-80-7
	SD	No	25-68-7	25-68-7
ANN	No	No	41-10-49	41-8-51
	MDL	No	42-7-51	43-6-51
	SD	No	37-8-55	37-7-56
1-NN	No	No	27-5-68	27-5-68
	MDL	No	35-8-57	36-7-57
	SD	No	33-6-61	33-6-61

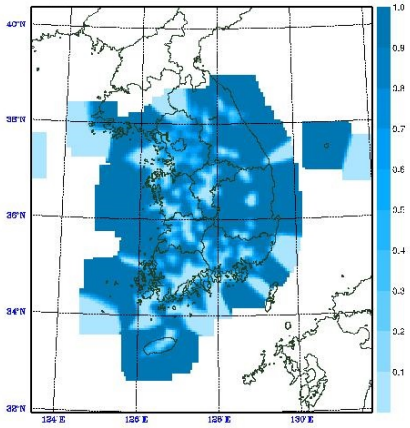
Tests were conducted individually on each station with the significance level at 0.01.
For ANN and 1-NN, the win-tie-loss percentage is calculated on the result of a single run.
Each item indicates the win-tie-loss percentage of the EWS model that used the PCA comparing with the original model in that row.



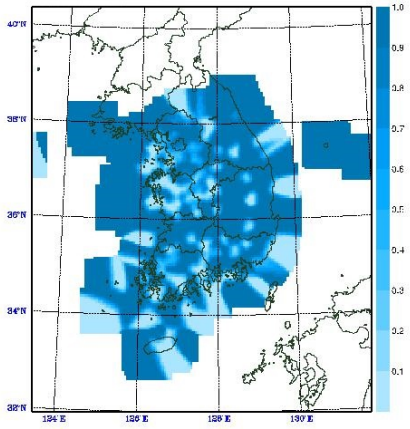
(a) Selective discretization without PCA



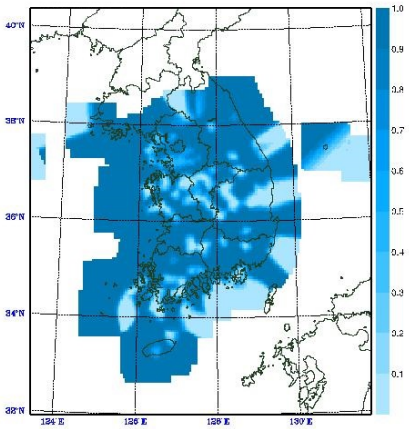
(b) Selective discretization with PCA



(c) PCA without selective discretization



(d) PCA with selective discretization



(e) Selective discretization and PCA

Figure 4.7: Effects of the preprocessing methods on logistic regression

Each figure shows the effect of the preprocessing method. Dark regions indicate areas where ETS was improved by the preprocessing method, while bright ones indicate areas where ETS was lowered.

Table 4.10: Wilcoxon signed-rank tests whether or not the selective discretization coupled with PCA significantly improves the predictive performance

Classifier	Discretization	PCA	F-measure	ETS
Logistic	No	No	61-31-8	61-31-8
C4.5	No	No	23-60-17	23-59-18
Forests	No	No	27-57-16	28-56-16
LIBSVM	No	No	48-37-15	49-36-15
SMO	No	No	79-21-0	87-12-0
RIPPER	No	No	31-61-8	31-59-10
ANN	No	No	58-8-34	57-8-35
1-NN	No	No	54-4-42	54-4-42

Tests were conducted individually on each station with the significance level at 0.01.
For ANN and 1-NN, the win-tie-loss percentage is calculated on the result of a single run.
Each item indicates the win-tie-loss percentage of the EWS model that used the both preprocessing methods comparing with the original model in that row.

since the nominal-to-binary filter of the PCA increased the number of input variables, which resulted in long preprocessing time; on the other hand, the selective discretization matched well with the PCA because they provided a good trade-off between the computation time and the predictive accuracy. One can see that the proposed method, which is the logistic regression with the selective discretization and the PCA, runs at a reasonable speed.

4.3.4 Discussions

In this section, various machine learning techniques were applied to the EWS for heavy rainfall nowcasting. The EWS that uses the logistic regression with selective discretization and PCA was proposed. The selective discretization method selectively discretized input variables that have a nonlinear relationship with the very short-range heavy rainfall, and the PCA reduced the dimensionality of input variables by creating a new coordinate system that provides an informative view of the data. The preprocessing methods could improve the prediction quality of the EWS further when used together than used separately. Empirical

Table 4.11: Running time analysis of various EWS models

Classifier	Discretization	PCA	Preprocessing	Training	Testing	Total
Logistic	No	No	0	7,423	236	7,659
	MDL	No	1,256	32,744	301	34,301
	SD	No	1,287	11,315	252	12,854
	No	Yes	2,330	3,870	226	6,426
	MDL	Yes	14,589	16,663	248	31,501
	SD	Yes	4,855	6,440	227	11,522
C4.5	No	No	0	5,433	25	5,458
	MDL	No	1,256	525	27	1,808
	SD	No	1,287	5,801	26	7,114
	No	Yes	2,330	2,972	23	5,325
	MDL	Yes	14,589	9,706	27	24,322
	SD	Yes	4,855	5,648	28	10,531
Forests	No	No	0	8,673	108	8,781
	MDL	No	1,256	1,342	106	2,704
	SD	No	1,287	7,657	111	9,055
	No	Yes	2,330	15,971	137	18,438
	MDL	Yes	14,589	13,049	132	27,770
	SD	Yes	4,855	16,340	135	21,330
LIBSVM	No	No	0	20,550	5,388	25,938
	MDL	No	1,256	17,341	3,316	21,913
	SD	No	1,287	22,016	5,056	28,359
	No	Yes	2,330	19,270	4,838	26,438
	MDL	Yes	14,589	41,780	9,999	66,368
	SD	Yes	4,855	27,645	6,608	39,108
SMO	No	No	0	1,048	244	1,292
	MDL	No	1,256	3,532	371	5,159
	SD	No	1,287	1,667	300	3,254
	No	Yes	2,330	13,115	79	15,524
	MDL	Yes	14,589	26,978	91	41,658
	SD	Yes	4,855	10,972	86	15,913
RIPPER	No	No	0	20,633	25	20,658
	MDL	No	1,256	6,082	25	7,363
	SD	No	1,287	18,918	26	20,231
	No	Yes	2,330	13,798	25	16,153
	MDL	Yes	14,589	19,686	25	34,300
	SD	Yes	4,855	17,203	25	22,083
ANN	No	No	0	198,858	1,896	200,754
	MDL	No	1,256	1,345,247	12,639	1,359,142
	SD	No	1,287	327,686	3,068	332,041
	No	Yes	2,330	59,253	587	62,170
	MDL	Yes	14,589	233,717	2,199	250,505
	SD	Yes	4,855	87,020	847	92,722
1-NN	No	No	0	38	218,290	218,328
	MDL	No	1,256	42	616,216	617,514
	SD	No	1,287	38	345,446	346,771
	No	Yes	2,330	36	294,308	251,674
	MDL	Yes	14,589	40	601,228	615,857
	SD	Yes	4,855	35	352,586	357,476

Experiments were conducted on the i5-760 processor with a clock rate of 2.8 GHz. Elapsed time for each station was measured in milliseconds and averaged over all stations.

results indicated that the proposed method works well on the very short-range heavy rainfall prediction in terms of F-measure and ETS.

It may be more natural to use regression functions than to use classifiers in predicting heavy rainfall, and the regression functions showed promising results in our preliminary work; however, most of the regression functions we tried were slow compared to the classifiers we used in this section. Our future work aims to further improve both the computation time and the prediction quality of the EWSs that use regression functions.

Chapter 5

Conclusions

The accuracy of short-range weather forecasts in South Korea were improved one step further by introducing the regional weather forecast system at the village level in 2008. However, the accuracy of short-term weather forecasts has not improved significantly since then due to the limitations of the numerical weather prediction (NWP). This thesis investigated the possibility of employing machine learning techniques to complement NWP on small spatial and temporal scales.

In Chapter 2, we introduced two dimensional reduction techniques: feature extraction and feature selection. The dimensional reduction techniques transformed the input variables or reduced the number of input variables to overcome the curse of dimensionality. It is recommended to use feature selection when various types of meteorological variables are used as input, and feature extraction when highly correlated variables are used. We also proposed a scheme for precipitation type predictions using numerous weather variables. Correlation-based feature selection was used to choose a discriminatory subset of input variables, and multinomial logistic regression was applied to the selected variables to predict wintertime precipitation types. Through feature selection, we successfully improved the accuracy of precipitation type forecasts from European Centre for Medium-Range Weather Forecast

and Regional Data Assimilation and Prediction System by more than 13 %.

In Chapter 3, we suggested sampling techniques to alleviate the class imbalance problem, which arise when predicting rare meteorological events. To balance the class distribution, undersampling removes common instances, while oversampling adds rare ones. Undersampling is preferred when the size of the original dataset is too large, and oversampling is preferred when there are not enough rare instances. We also proposed a machine learning approach to predict lightning within a particular location and time interval. We used an undersampling technique to overcome the class imbalance problem and to speed up the training process. We then trained support vector machines (SVMs) to forecast lightning. When trained with the original dataset, SVMs could not predict any lightning. After undersampling, however, lightning was successfully predicted. We further improved the performance of SVMs by extending the temporal and spatial scales.

In Chapter 4, we presented selective discretization scheme to selectively discretize input variables. Discretization is the process of converting continuous variables to categorical ones. Many learning algorithms can benefit from the discretization due to the enhanced learning speed and accuracy, however, information loss is inevitable. The selective discretization scheme can prevent information loss caused by inappropriate discretization. Selective discretization can be used when it is not easy to determine whether or not a particular continuous variable used as input is suitable for machine learning. We also proposed an early warning system (EWS) for very short-term heavy rainfall. The selective discretization method selectively discretized input variables such as date and temperature, and principal component analysis reduced the dimension of input variables. The predictive performance of the EWS was significantly improved by the two preprocessing methods. Empirical results showed that the proposed method worked well on the very short-term heavy rainfall prediction in terms of F-measure and ETS.

We suggested appropriate techniques for various problems that can be encountered when

performing meteorological forecast with machine learning. While our applications of machine learning to short-range meteorological forecasts were quite effective for the selected problems, there are still a wide variety of problems to be investigated. For example, regression is more appropriate than classification for rainfall or lightning forecasts. However, regression functions were very slow and produced unsatisfactory results compared with classifiers. It remains for further study to applying regression functions to short-range meteorological forecasts.

Bibliography

Abe, S., 2010. Feature selection and extraction. In: Support Vector Machines for Pattern Classification, pp. 331–341. Springer, London.

Aha, D. and Kibler, D., 1991. Instance-based learning algorithms. Mach. Learn. 6, 37–66.

Alfieri, L., Salamon, P., Pappenberger, F., Wetterhall, F., and Thielen, J., 2012. Operational early warning systems for water-related hazards in Europe. Environ. Sci. Policy 21, 35–49.

Allen, D.J. and Pickering, K.E., 2002. Evaluation of lightning flash rate parameterizations for use in a global chemical transport model. J. Geophys. Res. 107 (D23), 4711.

Behrangi, A., Yin, X., Rajagopal, S., Stampoulis, D., and Ye, H., 2018. On distinguishing snowfall from rainfall using near-surface atmospheric information: comparative analysis, uncertainties and hydrologic importance. Q. J. Roy. Meteor. Soc. 144 (S1), 89–102.

Bergmeir, C. and Benitez, J.M., 2012. On the use of cross-validation for time series predictor evaluation. Inform. Sci. 191, 192–213.

Boser, B.E., Guyon, I.M., and Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT), pp. 144–152 New York, NY, USA.

Box, J.E., Fettweis, X., Stroeve, J.C., Tedesco, M., Hall, D.K., and Steffen, K., 2012. Greenland ice sheet albedo feedback: thermodynamics and atmospheric drivers. Cryosphere 6 (4), 821–839.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Bright, D.R., Wandishin, M.S., Jewell, R.E., and Weiss, S.J., 2005. A physically based parameter for lightning prediction and its calibration in ensemble forecasts. Conference on Meteorological Applications of Lightning Data, San Diego, CA, USA.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Disc. 2 (2), 121–167.

Cardie, C. and Howe, N., 1997. Improving minority class prediction using case-specific feature weights. In: Proceedings of the Fourteenth International Conference on Machine Learning (ICML), pp. 57–65.

Caruana, R. and Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms. In: Proceedings of the Twenty-third International Conference on Machine Learning (ICML), pp. 161–168.

Cavalcanti, G.D., Ren, T.I., and Pereira, J.F., 2013. Weighted modular image principal component analysis for face recognition. Expert Syst. Appl. 40 (12), 4971–4977.

Cessie, S.L. and Houwelingen, J.C.V., 1992. Ridge estimators in logistic regression. J. Roy. Stat. Soc. C Appl. Stat. 41 (1), 191–201.

Chang, C.C. and Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2 (3), 1–27.

Chang, F.J., Chen, P.A., Lu, Y.R., Huang, E., and Chang, K.Y., 2014. Real-time multi-step-ahead water level forecasting by recurrent neural networks for urban flood control. J. Hydrol. 517, 836–846.

Chang, F.J. and Tsai, M.J., 2016. A nonlinear spatio-temporal lumping of radar rainfall for modeling multi-step-ahead inflow forecasts by data-driven techniques. J. Hydrol. 535, 256–269.

Chang, L.C., Amin, M., Yang, S.N., and Chang, F.J., 2018. Building ANN-based regional multi-step-ahead flood inundation forecast models. Water 10 (9), 1283.

Chattopadhyay, S. and Chattopadhyay, G., 2010. Univariate modelling of summer-monsoon rainfall time series: comparison between ARIMA and ARNN. C. R. Geosci. 342 (2), 100–107.

Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority oversampling technique. J. Artif. Intell. Res. 16, 321–357.

Chen, R.S., Liu, J.F., and Song, Y.X., 2014. Precipitation type estimation and validation in China. J. Mt. Sci. 11 (4), 917–925.

Clark, A.J., Gallus, W.A., and Weisman, M.L., 2010. Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF model simulations and the operational NAM. Weather Forecast. 25 (5), 1495–1509.

Cohen, W.W., 1995. Fast effective rule induction. In: Proceedings of the Twelfth International Conference on Machine Learning (ICML), pp. 115–123.

Cortes, C. and Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

Cover, T.M. and Hart, P.E., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theor. 13 (1), 21–27.

Cover, T.M. and Thomas, J.A., 2006. Elements of Information Theory. Wiley-Interscience.

Curran, E.B., Holle, R.L., and Lopez, R.E., 2000. Lightning casualties and damages in the United States from 1959 to 1994. J. Climate 13 (19), 3448–3464.

Dafis, S., Fierro, A., Giannaros, T.M., Kotroni, V., Lagouvardos, K., and Mansell, E., 2018. Performance evaluation of an explicit lightning forecasting system. J. Geophys. Res-Atmos. 123 (10), 5130–5148.

Dai, A., 2008. Temperature and pressure dependence of the rain-snow phase transition over land and ocean. Geophys. Res. Lett. 35 (12).

Davis, J. and Goadrich, M., 2006. The relationship between Precision-Recall and ROC curves. In: Proceedings of the Twenty-third International Conference on Machine Learning (ICML), pp. 233–240.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30.

Ding, B., Yang, K., Qin, J., Wang, L., Chen, Y., and He, X., 2014. The dependence of precipitation types on surface elevation and meteorological conditions and its parameterization. J. Hydrol. 513, 154–163.

Dougherty, J., Kohavi, R., and Sahami, M., 1995. Supervised and unsupervised discretization of continuous features. In: Proceedings of the Twelfth International Conference on Machine Learning (ICML), pp. 194–202.

Dubey, R., Zhou, J., Wang, Y., Thompson, P., and Ye, J., 2014. Analysis of sampling techniques for imbalanced data: an n=648 ADNI study. Neuroimage 87, 220–241.

Estabrooks, A., Jo, T., and Japkowicz, N., 2004. A multiple resampling method for learning from imbalanced data sets. Comput. Intell. 20 (1), 18–36.

Estévez, P.A., Tesmer, M., Perez, C.A., and Zurada, J.M., 2009. Normalized mutual information feature selection. IEEE T. Neural. Networ. 20 (2), 189–201.

Fan, R.E., Chen, P.H., and Lin, C.J., 2005. Working set selection using second order information for training support vector machines. J. Mach. Learn. Res. 6, 1889–1918.

Fayyad, U.M. and Irani, K.B., 1993. Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI), pp. 1022–1027.

Fierro, A.O., Gao, J., Ziegler, C.L., Mansell, E.R., MacGorman, D.R., and Dembek, S.R., 2014. Evaluation of a cloud-scale lightning data assimilation technique and a 3DVAR method for the analysis and short-term forecast of the 29 June 2012 derecho event. Mon. Weather Rev. 142 (1), 183–202.

Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3, 1289–1305.

Froidurot, S., Zin, I., Hingray, B., and Gautheron, A., 2014. Sensitivity of precipitation phase over the Swiss Alps to different meteorological variables. J. Hydrometeorol. 15 (2), 685–696.

Gao, X., Ye, B., Zhang, S., Qiao, C., and Zhang, X., 2010. Glacier runoff variation and its influence on river runoff during 1961–2006 in the Tarim river basin, China. Sci. China Earth Sci. 53 (6), 880–891.

Giannaros, T.M., Kotroni, V., and Lagouvardos, K., 2016. WRF-LTNGDA: a lightning data assimilation technique implemented in the WRF model for improving precipitation forecasts. Environ. Modell. Softw. 76, 54–68.

Giannaros, T.M., Lagouvardos, K., and Kotroni, V., 2017. Performance evaluation of an operational lightning forecasting system in Europe. Nat. Hazards 85 (1), 1–18.

Glossary of Meteorology, 2019a. Nowcast. Available from `http://glossary.ametsoc.org/wiki/Nowcast`.

Glossary of Meteorology, 2019b. Very short-range forecast. Available from `http://glossary.ametsoc.org/wiki/Very_short-range_forecast`.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H., 2009. The WEKA data mining software: an update. SIGKDD Explor. 11 (1), 10–18.

Hall, M.A., 1999. Correlation-based Feature Selection for Machine Learning. PhD thesis University of Waikato.

He, H. and Garcia, E.A., 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21 (9), 1263–1284.

Heffer, C.O., 2013. ELLA policy brief: Rio de Janeiro city's early warning system for heavy rain. ELLA, Pracitical Action Consulting, Lima, Peru.

Holle, R., 2008. Annual rates of lightning fatalities by country. Twentieth International Lightning Detection Conference (ILDC), Tucson, AZ, USA.

Hong, W.C., 2008. Rainfall forecasting by technological machine learning models. Appl. Math. Comput. 200 (1), 41–57.

Hornik, K., 1991. Approximation capabilities of multilayer feedforward networks. Neural Networks 4 (2), 251–257.

Hosmer, D.W., Lemeshow, S., and Sturdivant, R.X., 2013. Applied Logistic Regression. Wiley-Interscience.

Hsu, N.S., Huang, C.L., and Wei, C.C., 2015. Multi-phase intelligent decision model for reservoir real-time flood control during typhoons. J. Hydrol. 522, 11–34.

Ivanciuc, O., 2007. Applications of support vector machines in chemistry. Rev. Comp. Ch. 23, 291–400.

Japkowicz, N., 2000. The class imbalance problem: significance and strategies. In: Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI), pp. 111–117.

Jennings, K.S., Winchell, T.S., Livneh, B., and Molotch, N.P., 2018. Spatial variation of the rain-snow temperature threshold across the Northern Hemisphere. Nat. Commun. 9 (1), 1148.

Jin, R., Breitbart, Y., and Muoh, C., 2009. Data discretization unification. Knowl. Inf. Syst. 19 (1), 1–29.

Jolliffe, I.T., 2002. Principal Component Analysis. Springer Verlag.

Jolliffe, I.T. and Stephenson, D.B., 2003. Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wiley.

Keeter, K.K. and Cline, J.W., 1991. The objective use of observed and forecast thickness values to predict precipitation type in North Carolina. Weather Forecast. 6 (4), 456–469.

Kienzle, S.W., 2008. A new temperature based method to separate rain and snow. Hydrol. Process. 22 (26), 5067–5085.

Kim, Y.H., Choi, D.Y., Chang, D.E., Yoo, H.D., and Jin, G.B., 2011. An improvement on the criteria of special weather report for heavy rain considering the possibility of rainfall damage and the recent meteorological characteristics. Atmosphere. 21 (4), 481–495.

Kim, Y.H. and Yoon, Y., 2016. Spatiotemporal pattern networks of heavy rain among automatic weather stations and very-short-term heavy-rain prediction. Adv. Meteorol. 2016.

Kobayashi, M., 2018. Early warning system for heavy rain bursts put to test in Tokyo. The Asahi Shimbun. 24 July, Available from `http://www.asahi.com/ajw/`.

Korea Meteorological Administration, 2018. Criteria for advisory/warning information. Available from `http://web.kma.go.kr/eng/weather/forecast/standard_warning_info.jsp`.

Kwak, N. and Choi, C.H., 2002. Input feature selection for classification problems. IEEE T. Neural. Networ. 13 (1), 143–159.

Lee, J., Kim, J., Lee, J.H., Cho, I.H., Lee, J.W., Park, K.H., and Park, J., 2012. Feature selection for heavy rain prediction using genetic algorithms. In: Proceedings of the Joint Sixth International Conference on Soft Computing and Intelligent Systems and the Thirteenth International Symposium on Advanced Intelligent Systems (SCIS-ISIS), pp. 830–833.

Lee, M.K., Moon, S.H., Yoon, Y., Kim, Y.H., and Moon, B.R., 2018. Detecting anomalies in meteorological data using support vector regression. Adv. Meteorol. 2018.

Lee, S.M., Han, S.U., Won, H.Y., Ha, J.C., Lee, Y.H., Lee, J.H., and Park, J.C., 2014. A method for the discrimination of precipitation type using thickness and improved Matsuo's scheme over South Korea. Atmosphere 24, 151–158.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., and Liu, H., 2017. Feature selection: a data perspective. ACM Comput. Surv. 50 (6), 94.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S., 1997. Development and test of the distributed HBV-96 hydrological model. J. Hydrol. 201 (1), 272–288.

Liu, H., Hussain, F., Tan, C., and Dash, M., 2002. Discretization: An enabling technique. Data Min. Knowl. Discov. 6 (4), 393–423.

Liu, J.N.K., Li, B.N.L., and Dillon, T.S., 2001. An improved naïve Bayesian classifier technique coupled with a novel input solution method. IEEE Trans. Syst. Man Cybern. C Appl. Rev. 31 (2), 249–256.

Liu, S., Yan, D., Qin, T., Weng, B., Lu, Y., Dong, G., and Gong, B., 2018. Precipitation phase separation schemes in the Naqu river basin, eastern Tibetan plateau. Theor. Appl. Climatol. 131 (1), 399–411.

Liu, W., Wang, S., Zhou, Y., Wang, L., Zhu, J., and Wang, F., 2016. Lightning-caused forest fire risk rating assessment based on case-based reasoning: a case study in DaXingAn Mountains of China. Nat. Hazards 81 (1), 347–363.

Liu, X.Y., Wu, J., and Zhou, Z.H., 2009. Exploratory undersampling for class-imbalance learning. IEEE Trans. Syst. Man Cybern. B Cybern. 39 (2), 539–550.

Lynn, B.H., Kelman, G., and Ellrod, G., 2015. An evaluation of the efficacy of using observed lightning to improve convective lightning forecasts. Weather Forecast. 30 (2), 405–423.

Mäkelä, A., Saltikoff, E., Julkunen, J., Juga, I., Gregow, E., and Niemela, S., 2013. Cold-season thunderstorms in Finland and their effect on aviation safety. B. Am. Meteorol. Soc. 94 (6), 847–858.

Mecklenburg, S., Joss, J., and Schmid, W., 2000. Improving the nowcasting of precipitation in an alpine region with an enhanced radar echo tracking algorithm. J. Hydrol. 239, 46–68.

Mein, R.G. and Larson, C.L., 1973. Modeling infiltration during a steady rain. Water Resour. Res. 9 (2), 384–394.

Meyer, H., Kuhnlein, M., Appelhans, T., and Nauss, T., 2016. Comparison of four machine learning algorithms for their applicability in satellite-based optical rainfall retrievals. Atmos. Res. 169, 424–433.

Moon, S.H., Kim, Y.H., Lee, Y.H., and Moon, B.R., 2019. Application of machine learning to an early warning system for very short-term heavy rainfall. J. Hydrol. 568, 1042–1054.

Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. MIT Press.

Nanda, T., Sahoo, B., Beria, H., and Chatterjee, C., 2016. A wavelet-based non-linear autoregressive with exogenous inputs (WNARX) dynamic neural network model for real-time flood forecasting using satellite-based rainfall products. J. Hydrol. 539, 57–73.

Nastos, P., Paliatsos, A., Koukouletsos, K., Larissi, I., and Moustris, K., 2014. Artificial neural networks modeling for forecasting the maximum daily total precipitation at Athens, Greece. Atmos. Res. 144, 141–150.

Nese, J.M., Grenci, L.M., and Babb, D., 2018. A World of Weather: Fundamentals of Meteorology. Kendall Hunt.

Norrman, J., Eriksson, M., and Lindqvist, S., 2000. Relationships between road slipperiness, traffic accident risk and winter road maintenance activity. Clim. Res. 15 (3), 185–193.

Peng, H., Long, F., and Ding, C., 2005. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE T. Pattern Anal. 27 (8), 1226–1238.

Platt, J., 1999. Fast training of support vector machines using sequential minimal optimization. In: Schölkopf, B., Burges, C.J.C., Smola, A.C. (Eds.), Advances in Kernel Methods. MIT Press, Cambridge.

Price, C. and Rind, D., 1992. A simple lightning parameterization for calculating global lightning distributions. J. Geophys. Res. 97 (D9), 9919–9933.

Qian, W., Jiang, N., and Du, J., 2016. Anomaly-based weather analysis versus traditional total-field-based weather analysis for depicting regional heavy rain events. Weather Forecast. 31 (1), 71–93.

Quinlan, R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann.

Ralph, F.M., Rauber, R.M., Jewett, B.F., Kingsmill, D.E., Pisano, P., Pugner, P., Rasmussen, R.M., Reynolds, D.W., Schlatter, T.W., Stewart, R.E., Tracton, S., and Waldstreicher, J.S., 2005. Improving short-term (0-48 h) cool-season quantitative precipitation forecasting: recommendations from a USWRP workshop. B. Am. Meteorol. Soc. 86 (11), 1619–1632.

Ramírez, M.C.V., de Campos Velho, H.F., and Ferreira, N.J., 2005. Artificial neural network technique for rainfall forecasting applied to the São Paulo region. J. Hydrol. 301, 146–162.

Reeves, H.D., Ryzhkov, A.V., and Krause, J., 2016. Discrimination between winter precipitation types based on spectral-bin microphysical modeling. J. Appl. Meteorol. Clim. 55 (8), 1747–1761.

Rich, E., Knight, K., and Shivashankar, N.B., 2009. Artificial Intelligence. McGraw-Hill Education.

Rivera, D., Lillo, M., Uvo, C.B., Billib, M., and Arumí, J.L., 2012. Forecasting monthly precipitation in Central Chile: a self-organizing map approach using filtered sea surface temperature. Theor. Appl. Climatol. 107 (1), 1–13.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J., 1986. Learning representations by back propagating errors. Nature 323, 533–536.

Schultz, C.J., Petersen, W.A., and Carey, L.D., 2009. Preliminary development and evaluation of lightning jump algorithms for the real-time detection of severe weather. J. Appl. Meteorol. and Clim. 48 (12), 2543–2563.

Seo, J.H., Lee, Y.H., and Kim, Y.H., 2014. Feature selection for very short-term heavy rainfall prediction using evolutionary computation. Adv. Meteorol. 2014.

Sims, E.M. and Liu, G., 2015. A parameterization of the probability of snow-rain transition. J. Hydrometeorol. 16 (4), 1466–1477.

Su, C.T., Chen, L.S., and Yih, Y., 2006. Knowledge acquisition through information granulation for imbalanced data. Expert Syst. Appl. 31 (3), 531–541.

Sun, Y., Kamel, M.S., Wong, A.K., and Wang, Y., 2007. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognit. 40 (12), 3358–3378.

Theodoridis, S. and Koutroumbas, K., 2008. Pattern Recognition. Academic Press.

Toth, E., Brath, A., and Montanari, A., 2000. Comparison of short-term rainfall prediction models for real-time flood forecasting. J. Hydrol. 239, 132–147.

Vapnik, V. and Lerner, A., 1963. Pattern recognition using generalized portrait method. Automat. Rem. Contr. 24 (6), 774–780.

Wigmosta, M.S., Vail, L.W., and Lettenmaier, D.P., 1994. A distributed hydrology-vegetation model for complex terrain. Water Resour. Res. 30 (6), 1665–1679.

Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences. Academic Press.

Witten, I.H., Frank, E., Hall, M.A., and Pal, C., 2016. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

Wolpert, D.H. and Macready, W.G., 1997. No free lunch theorems for optimization. IEEE Trans. Evol. Comput. 1 (1), 67–82.

World Meteorological Organization, 2019. Definitions of meteorological forecasting ranges. Available from `https://www.wmo.int/pages/prog/www/DPS/gdps.html`.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.H., Steinbach, M., Hand, D.J., and Steinberg, D., 2007. Top 10 algorithms in data mining. Knowl. Inf. Syst. 14 (1), 1–37.

Wunsch, A., Liesch, T., and Broda, S., 2018. Forecasting groundwater levels using nonlinear autoregressive networks with exogenous input (narx). J. Hydrol. 567, 743–758.

Yair, Y., Lynn, B., Price, C., Kotroni, V., Lagouvardos, K., Morin, E., Mugnai, A., and Llasat, M.d.C., 2010. Predicting the potential for lightning activity in mediterranean storms based on the weather research and forecasting (WRF) model dynamic and microphysical fields. J. Geophys. Res. 115, D04205.

Yang, R. and Ren, M., 2011. Wavelet denoising using principal component analysis. Expert Syst. Appl. 38 (1), 1073–1076.

Yang, Z.L., Dickinson, R.E., Robock, A., and Vinnikov, K.Y., 1997. Validation of the snow submodel of the biosphere-atmosphere transfer scheme with Russian snow cover and meteorological observational data. J. Climate 10 (2), 353–373.

Yaseen, Z.M., Ebtehaj, I., Bonakdari, H., Deo, R.C., Mehr, A.D., Mohtar, W.H.M.W., Diop, L., El-shafie, A., and Singh, V.P., 2017. Novel approach for streamflow forecasting using a hybrid ANFIS-FFA model. J. Hydrol. 554, 263–276.

Zhang, D., Lindholm, G., and Ratnaweera, H., 2018. Use long short-term memory to enhance Internet of Things for combined sewer overflow monitoring. J. Hydrol. 556, 409–418.

Zhong, K., Zheng, F., Xu, X., and Qin, C., 2018. Discriminating the precipitation phase based on different temperature thresholds in the Songhua river basin, China. Atmos. Res. 205, 48–59.

국문 초록

기계 학습은 주어진 데이터를 통해 자동으로 프로그램을 생성해내는 기법으로서 인공지능의 한 분야이다. 특정 업무를 수행하기 위해 일련의 구체적인 명령어를 직접 기입해야만 했던 종래의 프로그래밍과 구분되며, 자연어 처리나 컴퓨터 비전에서와 같이 효과적인 알고리즘을 개발하기 힘든 경우 기계 학습이 선호된다.

전통적으로 기상 예보는 수치 예보 기법을 통해 이루어진다. 수치 예보는 현재의 기상 정보를 바탕으로 수학적 모델을 이용한 시뮬레이션을 통해 미래의 날씨를 예측한다. 하지만 수치 예보 기법은 초기 자료로 사용한 데이터에 오류가 있을 경우 시뮬레이션을 해나가며 그 오류가 증폭되고, 시공간적으로 비교적 낮은 해상도를 지니고 있으며, 일정 시간이 지나야만 예보가 안정화되기 때문에 국소적이면서 단기적인 기상 예측 문제에는 적합하지 않다. 이를 해결하기 위해 주어진 예측 문제에 적절한 기계 학습 기법을 사용하여 효과적으로 단기 기상 예측을 수행하는 방법들을 제안한다.

첫 번째로, 고차원의 입력 데이터를 가지고 효과적인 예측 모델을 만들기 위한 차원 축소 기법들을 소개한다. 입력 데이터의 차원이 증가함에 따라 기계학습 기법들이 필요로 하는 시간이나 메모리 요구량이 폭발적으로 증가하는 차원의 저주가 발생하는데, 차원 축소 기술은 이를 완화하기 위한 기법들이다. 차원 축소 기술에는 특징 선택과 특징 추출이 있다. 특징 선택은 전체 입력 인자들 중 일부의 입력 인자들만을 선택하는 반면, 특징 추출은 고차원의 입력 데이터를 저차원의 공간에 투영한다. 상관 관계 기반의 특징 선택과 대표적인 특징 추출 기법인 주성분 분석이 제시되며, 차원 축소 기술을 사용한 기상 예측 사례로서 강수 유형 예측 모델이 제안된다. 해당 모델은 단기 기상 예보에 포함된 93개의 기상 인자를 입력으로 받아 겨울철 강수 유형을 예측한다. 유효한 입력 인자 집합을 선택하기 위해 특징 선택 기법을 사용하며, 다중 로지스틱 회귀는 선택된 입력 인자들을 이용하여 비, 눈, 그리고 진눈깨비 중 어느 형태로 강수가 발생할 것인지 예측하기 위해 사용된다. 본 예측 모델은 강수유형 예측 정확도를 13 % 이상 개선했으며, 본 모델에서 특징 선택은 통계적으로 유의한 수준으로 정확도를 향상시켰다.

두 번째로, 흔치 않은 기상 이벤트를 예측하는 데에 도움을 주는 샘플링 기법들이 소개된다. 훈련 데이터에 클래스가 불균형하게 분포하는 경우 기계 학습 기법들은 전체 정확도를 높이고자

희귀한 예제들에 대한 예측 성능을 희생하는 경향이 있다. 이러한 클래스 불균형 학습 문제를 해결하기 위해 언더샘플링 기법은 흔한 예제의 숫자를 줄인다. 언더샘플링 기법을 사용한 기상 예측 사례로서 뇌전 예측 모델이 제시된다. 해당 모델은 유럽 중기 예보 센터로부터 단기 기상 예보를 입력으로 받아 뇌전 발생 유무를 예측한다. 클래스 불균형 학습 문제를 해결하기 위해 언더샘플링이 사용되며, 지지 벡터 기계를 사용하여 특정 시간대에 특정 지역에서의 뇌전 발생 유무를 예측한다. 원래의 입력 데이터에서는 뇌전을 하나도 예측하지 못했지만 언더샘플링을 통해 약 38 %의 뇌전을 성공적으로 감지해냈다.

마지막으로, 이산화하기에 적합한 인자를 자동으로 선별하여 이산화하는 선택적 이산화 기법이 소개된다. 이산화는 연속형 변수를 범주형 변수로 변환하는 전처리 기법이다. 종래의 이산화 기법은 모든 변수에 대해 이산화를 적용하는데 이 과정에서 정보 손실은 불가피하다. 선택적 이산화 기법은 종속 변수와 비선형 관계에 있는 변수만을 이산화하여 정보 손실을 최소화한다. 이러한 선택적 이산화 기법을 사용한 기상 예측 사례로서 집중 호우 예측 모델이 제시된다. 본 모델은 자동 기상 관측 시스템으로부터 입력을 받아 세 시간 이내에 호우 주의보 조건이 충족될 것인지를 예측한다. 입력 데이터는 선택적 이산화 기법과 주성분 분석을 통해 응축된 양질의 정보를 담도록 전처리되고, 로지스틱 회귀는 전처리된 입력 데이터를 이용하여 호우 주의보 조건이 만족될 것인지 예측한다. 선택적 이산화 기법은 일자나 기온과 같은 인자들을 선택적으로 이산화하여 통계적으로 유의한 수준으로 예측 성능향상에 기여했다.

본 논문은 단기 기상 예보를 위한 효과적인 기계 학습 기법들을 제시하고, 강수 유형, 뇌전, 그리고 집중 호우 예측에 기계 학습을 효과적으로 적용한 사례들을 제공한다. 각 사례에서는 해당 예측 문제를 효과적으로 풀 수 있는 기법들을 조합했으며, 우리가 만든 예측 모델들은 실제 운용 목적으로 사용할 수 있을 정도의 성공적인 예측 품질을 보여주었다.

주요어 : 기계 학습, 기상 예측, 차원 축소, 언더샘플링, 이산화.

학번 : 2004-23567